Transitionen von der Erstausbildung ins Erwerbsleben
Transitions de l'Ecole à l'Emploi
Transitions from Education to Employment

TREE-LÄNGSSCHNITTGEWICHTUNG: KONSTRUKTION UND ANWENDUNG

Dokumentation zu den Erhebungswellen 2000 bis 2010

STEFAN SACCHI

TREE & cue sozialforschung, Basel & Zürich, 2011
www.tree.unibas.ch

Inhalt

Ausgangslage

Der Jugendlängsschnitt '*Transition from Education to Employment*' (TREE) beruht auf einer mehrwelligen Panelbefragung von Jugendlichen auf dem Weg von der Schule ins Berufsleben. Als Ausgangspunkt respektive *Basiserhebung* dient der schweizerische Teil der im Frühling 2000 durchgeführten *PISA-Studie* (PISA, 2002). Im Rahmen dieser international vergleichenden Erhebung werden Basiskompetenzen wie die Lesefähigkeit, aber auch das mathematisch-naturwissenschaftliche Grundwissen zusammen mit relevanten Kontext- und Hintergrundinformationen detailliert erhoben. Die von PISA befragten und getesteten Jugendlichen wurden 2001 bis 2010 im Rahmen des TREE-Projekts insgesamt acht mal nachbefragt – bis 2007 jeweils in jährlichem Abstand,– um so repräsentative Längsschnittinformationen insbesondere über die Entstehung und Bewältigung von Problemlagen beim Übergang von der Schule in die Berufsausbildung und später in den Arbeitsmarkt zu gewinnen.

Die vorliegende Dokumentation beschreibt die longitudinale Stichprobengewichtung für die Basisbefragung von PISA 2000 und die ersten acht Nachbefragungswellen des TREE-Panels. Es handelt sich um eine Aktualisierung der Dokumentation von 2008 (Sacchi 2008a) für die ersten sieben TREE-Wellen, die ihrerseits auf früheren Arbeitspapiere (Sacchi, 2003, 2004a, 2004b) aufbaut.

1 Stichprobe und Grundgesamtheit

Die als Basis für das TREE-Panel dienende PISA-Stichprobe ist so konzipiert, dass sie sowohl für Neuntklässler als auch – unabhängig von der aktuellen Klassenzuteilung – für Fünfzehnjährige repräsentativ ist. Es handelt sich um eine zweistufige, mehrfach disproportionale Zufallsauswahl mit vorgegebenem Stichprobenumfang für die beiden erwähnten Gruppen, für die Sprachregionen sowie für spezifische Kantone (für Details siehe Renaud, Ramseier & Zahner, 2000; PISA, 2002). In der Romandie ist zusätzlich und unabhängig davon eine Klassenstichprobe aus allen neunten Klassen gezogen worden, wobei jeweils alle Schüler und Schülerinnen in den ausgewählten Klassen befragt worden sind (einstufige Klumpenstichprobe, vgl. PISA Romandie, ohne Jahr).

Die TREE-Grundgesamtheit ist definiert durch die Teilmenge der PISA-Befragten, die zum Zeitpunkt der PISA-Erhebung eine öffentliche Regelschule irgendwo in der Schweiz oder eine private Regelschule in der italienischsprachigen Schweiz besucht und die ihre obligatorische Schulzeit damals noch nicht beendet hatten, die dann aber nach Ende des Schuljahres 1999/2000 aus der obligatorischen Schule ausgetreten sind. Die Untersuchungspopulation von TREE ist somit im Wesentlichen mit der PISA-Teilstichprobe von Schülern und Schülerinnen der neunten Klasse identisch.[1]

[1] Hinzu kommt eine kleine Gruppe von Jugendlichen der 15jährigen-Stichprobe, die zur Zeit der PISA-Erhebung eine siebte oder achte Klasse besucht und die die obligatorische Schule bereits im Schuljahr 1999/2000 vorzeitig verlassen haben (≈ 1 % der TREE-Ausgangsstichprobe).

Als Ausnahmen sind zu erwähnen, dass Schülerinnen und Schüler von Privatschulen lediglich in der italienischsprachigen Schweiz berücksichtigt, in den anderen beiden Landesteilen aber ausgeschlossen werden.[1] Zudem zählen auch jene PISA-Befragten, die ein Jahr *nach* der PISA-Basiserhebung noch nicht aus der obligatorischen Schule ausgetreten waren, nicht zur TREE-Grundgesamtheit.

2 Befragungsprozess und Beteiligung

Für die Panel-Befragungen im Rahmen der TREE-Längsschnitterhebung musste aus Datenschutzgründen schon im Rahmen der PISA-Erhebung das Einverständnis der befragten Schülerinnen und Schüler eingeholt werden. Den Jugendlichen ist zu diesem Zweck im Rahmen der PISA-Testsessions ein zielgruppengerechtes Informations-Blatt zum TREE-Projekt mit der Bitte, sich daran zu beteiligen, ausgehändigt worden. Zusätzlich sind auch die PISA-TestadministratorInnen mit einem eigenen Merkblatt instruiert worden (vgl. Meyer, 2000). Die teilnahmebereiten Jugendlichen sind gebeten worden, das mit ihrer Adresse versehene Informations-Blatt an TREE zurücksenden. Eine erste explorative Auswertung des resultierenden Adressblattrücklaufs lieferte dabei deutliche Hinweise, dass die *regionale bzw. lokale Testadministration* einen massgeblichen Einfluss auf die Beteiligung an der Adressenerhebung hatte (Meyer, 2000: 4). Jene Jugendlichen, die ihre Adresse zur Verfügung gestellt haben, sind erstmals im Frühling 2001 (Welle 1) und anschliessend in jährlichem Abstand jeweils im Frühling erneut befragt worden. *Tabelle 1* auf der folgenden Seite gibt einen Überblick über den Befragungsprozess bis zur achten und vorläufig letzten Folgebefragung im Jahr 2010.

Von den 7070 PISA-Befragten, die ihre Adresse für die Längsschnittstudie zur Verfügung gestellt haben, hat sich bei insgesamt 727 Fällen erst retrospektiv herausgestellt, dass sie nicht zur Grundgesamtheit gehören.[2] Die grosse Mehrheit der betreffenden Jugendlichen (N=608) hat im Jahr 2001 – bei der ersten TREE-Befragungswelle – entweder noch die obligatorische Schule besucht oder dann die neunte Klasse wiederholt.[3] Die übrigen gut 100 Fälle werden ausgeschlossen, weil sie zur Zeit der PISA-Erhebung keinen von TREE berücksichtigten Schultyp besucht haben (insb. Ausschluss von Sonderschulen). Die bereinigte Ausgangsstichprobe für die erste Welle des TREE-Panels umfasst somit noch 6343 Fälle. In *Tabelle 1* sind die Ausgangsstichprobe, die sich über die Befragungswellen kumulierenden Ausfälle (siehe dazu auch den Tabellenfuss) und der Rücklauf der insgesamt acht Befragungswellen bezogen auf die retrospektiv bereinigte Ausgangsstichprobe dokumentiert. Bei sämtlichen Wellen ist ein ausgezeichneter Rücklauf von 76 bis zu knapp 88 Prozent erreicht worden. In Kumulation resultiert so bis zur achten Welle eine für ein so langes und vielwelliges Panel sehr ansehnliche Beteiligung von knapp 54 Prozent. Dabei haben sich über 41 Prozent der Jugendlichen in der Ausgangsstichprobe an *sämtlichen* acht Befragungswellen beteiligt.[4]

---

[1] Die PISA-Bruttostichprobe (N=13467) reduziert sich dadurch um 673 Fälle bzw. um 5 Prozent.

[2] Die Bereinigung der Ausgangsstichprobe beruht dabei überwiegend auf den Angaben, die im Rahmen der ersten Welle erhoben worden sind, sowie in zweiter Linie auf indirekten Informationen über Nicht-Antwortende (Kontaktprotokolle).

[3] Der Befragungszeitpunkt für Welle 1 und 2 ist im Arbeitspapier von 2003 (Seite 2) falsch ausgewiesen; richtig ist Frühjahr 2001 respektive 2002 (zum Erhebungsdesign: Bundesamt für Statistik, 2003: 28-31).

[4] Die Angaben zur Beteiligung an den acht Befragungen entsprechen dem Datenstand vom Januar 2010, der auch der Stichprobengewichtung zugrunde liegt. Sie können geringfügig von späteren Werten abweichen.

Tabelle 1: *Ausgangsstichprobe, realisiertes Sample und Ausschöpfung*

	Nationale PISA-Stichprobe	Klassen-stichprobe Romandie	Kombinierte Ausgangs-Stichprobe	Realisiertes Sample (N)	Rück-lauf (%)
1. PISA-Erhebung					
Ausgangsstichprobe PISA	10423	?	?		
davon ausserhalb PISA-Grundgesamtheit [1)]	101	?	?		
davon abwesend	150	?	?		
PISA-Erhebung	10172	5073 [2)]	15241	14494	95.1 %
davon in beiden PISA-Stichproben				– 1031	
Kombiniertes PISA-Sample				13463	
davon ausserhalb TREE-Grundgesamtheit [3)]				– 673	
Bereinigte Ausgangsstichprobe Adressen-Erhebung				12794	
2. TREE-Panel					
Adressen-Erhebung			12794	7070	55.3 %
davon ausserhalb TREE-Grundgesamtheit [4)]				– 727	
bereinigte TREE-Ausgangsstichprobe				**6343**	
Befragungswelle 1			6343	5532	87.2 %
bei Welle 1 definitiv ausgeschieden [5)]			– 400		
Befragungswelle 2			5943	5210	87.7 %
bei Welle 2 definitiv ausgeschieden [6)]			– 344		
Befragungswelle 3			5599	4880	87.2 %
bei Welle 3 definitiv ausgeschieden [7)]			– 266		
Befragungswelle 4			5333	4680	87.8 %
bei Welle 4 definitiv ausgeschieden [8)]			– 284		
Befragungswelle 5			5049	4504	89.2 %
bei Welle 5 definitiv ausgeschieden [9)]			– 205		
Befragungswelle 6			4844	4135	85.4 %
bei Welle 6 definitiv ausgeschieden [10)]			– 204		
Befragungswelle 7			4640	3982	85,8 %
bei Welle 7 definitiv ausgeschieden [11)]			– 120		
bei Welle 8 irrtümlich nicht kontaktiert			– 15		
Befragungswelle 8			4505	3424	76,0 %
Kumulierte Ausschöpfung TREE T1 − T8				3982	54,0 %
davon Beteiligung an sämtlichen 8 Wellen				2618	41,3 %

1) Unfähig, die PISA-Testsession zu absolvieren. 2) Indirekt aus Stichprobengewichtung erschlossen. 3) Ausserhalb TREE-Grundgesamtheit: Jugendliche in Privatschulen (exkl. italienisches Sprachgebiet), 15jährige in nicht-obl. Ausbildung, Berner Jura. 4) Nach T2 retrospektiv ausgeschlossen, Details siehe Text. 5) Definitiv verweigert (N=204), Nicht auffindbar (N=195), Verstorben (N=1). 6) Definitiv verweigert (N=224), Nicht auffindbar (N=112), Ausgewandert (N=6), Verstorben (N=2). 7) Definitiv verweigert (N=251), Nicht auffindbar (N=3), Ausgewandert (N=11), Verstorben (N=1). 8) Definitiv verweigert, nicht mehr angeschrieben (N=266), Nicht auffindbar (N=7), Ausgewandert (N=9), Verstorben (N=2). 9) Definitiv verweigert, nicht weiter kontaktiert (N=189), nicht auffindbar (N=9), Ausgewandert (N=5), Verstorben (N=1). 10) Definitiv verweigert, nicht weiter kontaktiert (N=188), Nicht auffindbar (N=9), Ausgewandert (N=17), Verstorben (N=2). 11) Definitiv verweigert, nicht weiter kontaktiert (N=69), Nicht auffindbar (N=43), nachträglich als nicht zur Population gehörend ausgeschlossen (N=2) , Verstorben (N=6).

Die alles in allem aussergewöhnlich hohe Beteiligung verdankt sich nicht zuletzt dem Umstand, dass für potentielle Verweigerer jeweils alternative Beteiligungsformen vorgesehen worden sind. So hatten die Jugendlichen bei den ersten vier TREE-Wellen die Möglichkeit, die Fragen telefonisch anstatt schriftlich zu beantworten, wobei − wo nötig − auch ein erheb-

lich gekürzter Fragenkatalog eingesetzt worden ist (kurze Telefon-Interviews). Ab der fünften Befragungswelle kommt eine Kombination von computergestützten Telefoninterviews (CATI) und schriftlicher Befragung als Standardmodus der TREE-Erhebung zum Zug. Dieser *Methodenwechsel* ist auf der einen Seite angezeigt, weil die mit dem Alter zunehmende Vielfalt an individuellen Bildungs- und Erwerbsverläufen eine zunehmend komplexere, mit einer schriftlichen Befragung kaum noch vereinbare Filterführung erforderlich macht. Vor allem mit Blick auf die intraindividuelle Vergleichbarkeit einer Reihe von Angaben und psychologischen Skalen, die auf *Mode*-Effekte anfällig sein können (Klein & Porst; 2000; Scherpenzeel, 2001), ist es auf der anderen Seite aber auch wichtig, die betreffenden Informationen weiterhin schriftlich zu erheben.

Zwei Non-Response-Befragungen im Anschluss an die TREE-Erhebung 2003 und 2004 legen zudem ebenfalls einen Methodenwechsel nahe (Stalder & Dellenbach; 2005). Im Rahmen dieser zwei Zusatzerhebungen sind jeweils mehr als 1000 Jugendliche, die die Teilnahme an der betreffenden Erhebungswelle ganz verweigert oder sich nur telefonisch beteiligt haben, nach ihren Verweigerungsgründen gefragt haben. Gemäss diesen Non-Response-Analysen sind es vor allem Zeitgründe, die die Befragten dazu bewegen, sich nicht an der schriftlichen Befragung zu beteiligen oder die Teilnahme ganz zu verweigern. Jugendliche, die sich telefonisch (Kurz- oder Langinterview) beteiligt haben, kritisieren zudem die Länge und die zunehmende Komplexität des schriftlichen Erhebungsinstruments. Vollständige Verweigerungen werden jedoch eher mit mangelndem Interesse am Thema und dem Unwillen, sich weiterhin regelmässig an der Studie zu beteiligen, begründet. Diese Rückmeldungen stützen ebenfalls die Argumente zugunsten eines Methodenwechsels, der den Fokus auf ein im Vergleich zur Langversion des des alternativen Telefon-Interviews ein drastisch verkürztes CATI-Interview legt.

Sowohl theoretische wie auch empirische Gründe sprechen somit für den mit der *fünften Welle* umgesetzten Wechsel zu einem *Mixed-Mode-Design*, welches ein relativ kurzes Telefoninterview mit einem schriftlichen Ergänzungsfragebogen kombiniert. Alternativ zu diesem gemischten Standard-Modus kann zudem ab Welle 5 entweder das gesamte Befragungsprogramm schriftlich beantwortet werden, oder dann ein erheblich reduzierter Fragenkatalog telefonisch *oder* schriftlich (überwiegend CATI, vereinzelt schriftlich).

Wie aus *Tabelle 2* hervorgeht, hat sich dabei der Anteil der Antwortenden, der lediglich zu einer erheblich reduzierten Befragung bereit war, ab der dritten Welle sprunghaft erhöht. Ab T3 erhöht er sich zunächst von zwei bis drei auf fünfzehn Prozent, um dann ab der fünften Welle rund zwanzig Prozent zu erreichen. Die Möglichkeit einer hinsichtlich Umfang reduzierten Teilnahme (Alternativ-Modus 2) hat offenbar ab Welle 3 wesentlich zur erfreulichen Ausschöpfung beigetragen. Daneben dürfte sich diese auch dem konsequent eingesetzten Mixed-Mode-Design verdanken, das auch schriftlich weniger versierten Jugendlichen eine Teilnahme erheblich erleichtert hat.

Bereits aus Tabelle 1 geht hervor, dass ein sehr bedeutender Teil des Stichprobenschwunds bereits vor der ersten TREE-Befragung bei der Adressen-Erhebung im Rahmen der PISA-Testadministration entstanden ist. Dies verdeutlicht die folgende Grafik, die ersichtlich macht, wie sich der Umfang der jeweils verbleibenden Ausgangsstichprobe über die einzelnen Befragungsschritte hinweg sukzessive vermindert.[1]

[1] Der dargestellte Stichprobenschwund bezieht sich auf das *definitive* Ausscheiden aus der Panelstichprobe.

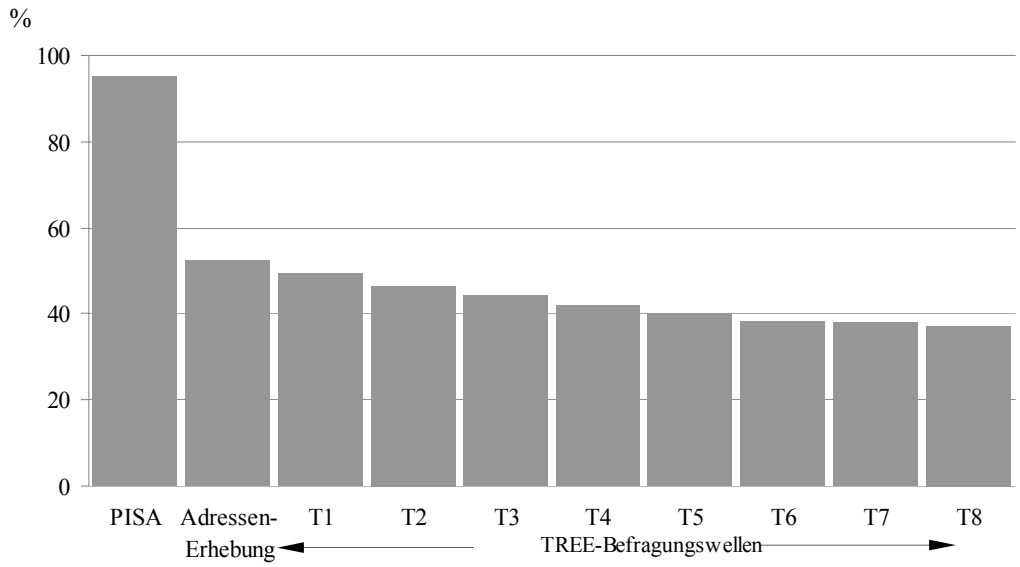
Tabelle 2: *Art der Teilnahme nach Befragungswelle*

Befragungsmethode (Anteil Teilnehmende)	Standard-Modus	Alternativ-Modus 1: volles Frageprogramm	Alternativ-Modus 2: reduz. Frageprogramm
TREE-Befragungswelle			
Welle 1 (N = 5532)	Schriftlich (91,3 %)	Telefoninterview lang (6,6 %)	Telefoninterview kurz (2,2 %)
Welle 2 (N = 5210)	Schriftlich (91,7 %)	Telefoninterview lang (5,0%)	Telefoninterview kurz (3,3 %)
Welle 3 (N = 4880)	Schriftlich (81,7 %)	Telefoninterview lang (3,0 %)	Telefoninterview kurz (15,4 %)
Welle 4 (N = 4680)	Schriftlich (81,3 %)	Telefoninterview lang (5,2 %)	Telefoninterview kurz (13,5 %)
Welle 5 (N = 4504)	CATI plus Schriftlich (76,5 %) [1)]	Schriftlich (3,1 %) [2)]	CATI (20,5 %) [3)]
Welle 6 (N = 4135)	CATI plus Schriftlich (79,7 %) [1)]	Schriftlich (1,1 %) [2)]	CATI (19,3 %) [3)]
Welle 7 (N = 3982)	CATI plus Schriftlich (74,0 %) [1)]	Schriftlich (5,7 %) [2)]	CATI (20,3 %) [3)]
Welle 8 (N = 3424)	CATI plus Schriftlich (71,6 %) [1)]	Schriftlich (9,1 %) [2)]	CATI (19,3 %) [3)]

1) Ergänzungsfragebogen insb. mit diversen für 'Mode'-Effekte anfälligen psychologischen Skalen. Anteil inklusive Einzelfälle mit abgebrochenem CATI-Interview. 2) Basis- (anstelle CATI-Interview) plus Ergänzungsfragebogen. 3) Ergänzungsfragebogen nicht ausgefüllt; enthalten sind auch einzelne Fälle mit ausgefülltem Basisfragebogen anstatt des CATI.

Da der Stichprobenschwund bzw. Nonresponse nur in seltenen Fällen als stichproben-neutral anzusehen ist (Schnell, 1997), wird dessen potenziell verzerrende Wirkung auf die Zusammensetzung des Samples im Rahmen von Panelstudien üblicherweise mit einer entsprechenden Längsschnittgewichtung kompensiert (siehe z. B. den Methodenvergleich von Rizzo, Kalton & Brick, 1994). Mit Blick auf TREE sollte ein Haupteffort dabei dem quantitativ weitaus gravierendsten Nonresponse gelten, der − noch vor der ersten TREE-Intervention − bei der Erhebung der Adressen von teilnahmewilligen Jugendlichen durch die PISA-Testadministration entstanden ist (siehe *Grafik 1*).

Grafik 1: *Kumulative Wirkung des Stichprobenschwunds (Ausgangsstichprobe = 100%)* [1]



[1] Anteil ProbandInnen, die bis zur jeweiligen Welle in der Panelstichprobe verbleiben (d. h. eine weitere Teilnahme nicht definitiv verweigern).

Bei einer Panel-Befragung bestehen allgemein bessere Möglichkeiten zur Korrektur von Nonresponse-Verzerrungen als bei einer einmaligen Querschnittbefragung, weil aus früheren Befragungswellen sehr viel reichhaltigere Informationen über die Nicht-Antwortenden vorliegen. Mit Blick auf die TREE-Gewichtung ist es dabei äusserst günstig, dass der Nonresponse bei der Erstbefragung im Rahmen von PISA mit zirka 5 Prozent minimal geblieben ist. Sämtliche im Rahmen von PISA erhobenen Informationen über die Befragten und den Befragungskontext können somit für die Korrektur des speziell beim Adress-Rücklauf substanziellen Nonresponse herangezogen werden. Zur Korrektur des bei den weiteren Wellen entstehenden Nonresponse können zusätzlich jeweils auch Informationen aus vorangehenden TREE-Wellen berücksichtigt werden.

3 Longitudinale Panel-Gewichtung für TREE

Panel-Gewichtungen für Personenstichproben werden gewöhnlich als Kehrwert des Produkts der Antwortwahrscheinlichkeiten der einzelnen Befragungswellen konstruiert (vgl. Sacchi, 2001), so etwa auch vom deutschen Sozio-ökonomischen Panel (Haisken-DeNew & Frick, 2000, 140f.). Bezogen auf das TREE-Panel ergibt sich somit:

$$G_i = \frac{1}{E_{PISA.i} \cdot A_{PISA.i}} \cdot \frac{1}{A_{ADR.i}} \cdot \frac{1}{A_{W1.i}} \cdot \ldots \cdot \frac{1}{A_{Wt.i}} \qquad (1)$$

wobei:

G_i	Longitudinales Gewicht zu Befragungswelle *t* für ProbandIn i
$E_{PISA.i}$	Einschlusswahrscheinlichkeit von *i* in die PISA-Ausgangstichprobe
$A_{PISA.i}$	Teilnahmewahrscheinlichkeit von *i* bei der PISA-Befragung
$A_{ADR.i}$	Teilnahmewahrscheinlichkeit von *i* bei der Adressen-Erhebung für TREE
$A_{W1.i}$	Teilnahmewahrscheinlichkeit von *i* bei TREE-Befragungswelle 1
$A_{Wt.i}$	Teilnahmewahrscheinlichkeit von *i* bei TREE-Befragungswelle *t*

Bei den Teilnahmewahrscheinlichkeiten $A_{..i}$ handelt es sich dabei jeweils um bedingte Wahrscheinlichkeiten, d. h. um die Antwortwahrscheinlichkeit unter der Voraussetzung, dass ProbandIn *i* zunächst in die PISA-Ausgangsstichprobe gelangt und dann nicht in einer der folgenden Befragungswellen wieder aus dem Panelstichprobe ausgeschieden ist. Der quantitativ weitaus wichtigste Grund für das Ausscheiden aus der Panelstichprobe ist dabei die *definitive* Verweigerung der Mitwirkung nicht nur an der aktuellen, sondern an *sämtlichen* noch ausstehenden Befragungswellen (siehe Tabelle 1, Tabellenfuss).[1] Hinzu kommt pro Welle jeweils eine Handvoll Jugendlicher, die ausgewandert oder verstorben sind, und die daher aus der Untersuchungspopulation ausscheiden, sowie eine ebenfalls meist kleine Anzahl, die weggezogen und nicht mehr auffindbar ist.

[1] Hingegen führt eine punktuelle Nicht-Teilnahme an einzelnen Befragungswellen nicht zum Ausscheiden aus dem Panel.

Ein Gewichtungsmodell gemäss Beziehung (1) hat mit Blick auf das TREE-Panel folgende Vorzüge:

- Das Modell baut auf der bestehenden PISA-Gewichtung auf, die Designeffekte und Nonresponse für das PISA-Sample korrigiert: Diese entspricht jeweils dem Kehrwert des Produkts von $E_{PISA.i}$ und $A_{PISA.i}$ gemäss Beziehung (1). Die übrigen Bausteine von G_i können – wie im Rahmen von Panelgewichtungen üblich – mittels Logit-Modellen respektive logistischen Regressionen bestimmt werden. Dies erlaubt es, *systematische* inter-individuelle Differenzen in den Teilnahmewahrscheinlichkeiten zu berücksichtigen.
- Die vielleicht nicht unbedingt zwingende Separierung von $A_{ADR.i}$ und $A_{W1.i}$ im Modell ermöglicht es zum einen, das Schwergewicht bei der Modellentwicklung auf die Schätzung von $A_{ADR.i}$ zu legen, auf das sich der Nonresponse und die damit verbundenen potentiellen Verzerrungen konzentrieren. Zum anderen trägt sie auch dem Umstand Rechnung, dass teilweise unterschiedliche Bestimmungsfaktoren für $A_{ADR.i}$ und $A_{W1.i}$ massgeblich sein dürften: Die Beantwortung des Antwortblattes dürfte wie erwähnt stark vom situativen Kontext der PISA-Testsessions beeinflusst sein, während für die Teilnahme an den weiteren Befragungswellen vor allem Individualmerkmale entscheidend sein werden.
- Dank dem baukastenartigen Aufbau kann die Gewichtung bei Bedarf problemlos um weitere Befragungswellen und entsprechende Nonresponse-Korrekturen erweitert werden, indem G_i gemäss Beziehung (1) mit dem Kehrwert der Antwortwahrscheinlichkeit der weiteren Welle(n) multipliziert wird.

Die folgenden Unterabschnitte (3.1-3.10) beschreiben zuerst schrittweise, wie die einzelnen *Bausteine des Gewichtungsmodells* gemäss Beziehung (1) berechnet werden. Darauf folgt eine zusammenfassende Betrachtung zu den *kumulativen Wirkungen des Nonresponse* auf die Zusammensetzung des Panelstichprobe (3.11). In den weiteren Abschnitten werden dann die *Stutzung* (Abschnitt 4) und *nachträgliche Schichtung* (Abschnitt 5) der resultierenden Rohgewichte sowie die verfügbaren *Typen von Gewichtungsvariablen* (Abschnitt 6) beschrieben. Ganz zum Schluss werden noch kurz ausgewählte forschungspraktische Aspekte mit Blick auf die *Anwendung der Gewichte* aufgegriffen (Abschnitt 7).

3.1 Konstruktion der Basisgewichtung für das zusammengesetzte PISA-Sample

Das PISA-Ausgangssample setzt sich wie erwähnt aus zwei *unabhängigen* Teilstichproben zusammen, für die je eine Gewichtungsvariable verfügbar ist, welche vor allem die aufgrund des Stichprobendesigns ungleichen Auswahlwahrscheinlichkeiten, aber auch den Nonresponse kompensiert. Die Nonresponse-Korrekturen beschränken sich auf den Ausgleich der schul- (nationale Stichprobe) bzw. klassenspezifischen (Klassensample Romandie) Ausfallquoten (vgl. auch Tabelle 1). Es handelt sich also – wie im Rahmen von Querschnitt-Erhebungen üblich – um eine nur sehr rudimentäre und nicht ganz unproblematische Korrektur, die aber angesichts eines Nonresponse von bloss ungefähr fünf Prozent auch kaum ins Gewicht fällt.[1] Die Konstruktion der beiden PISA-Gewichtungen ist an anderer Stelle dokumentiert (Renaud, Ramseier & Zahner, 2000; PISA Consortium, 2000b; PISA Romandie, ohne Jahr). Wichtig ist, dass es sich in beiden Fällen um eine Gewichtung nach dem Prinzip der reziproken Auswahlwahrscheinlichkeit handelt, welche sich in Beziehung (1) einsetzen lässt.

[1] Die PISA-Nonresponse-Korrekturen beruhen auf der nicht ganz unproblematischen Annahme, dass die Antwortenden einer Schule (nationale Stichprobe) bzw. Klasse (Klassensample Romandie) jeweils auch für die Nicht-Antwortenden repräsentativ sind.

Für die Berechnung der TREE-Gewichtung gemäss Beziehung (1) wird allerdings eine PISA-Basisgewichtung benötigt, welche Designeffekte und Nonresponse für das *zusammengesetzte PISA-Sample* korrigiert, das sowohl die nationale PISA-Stichprobe wie auch die − davon unabhängige − Klassenstichprobe für die Romandie umfasst. Für die deutsch- und italienischsprachige Schweiz, wo bloss eine Stichprobe gezogen wurde, kann dafür ohne weiteres die Gewichtungsvariable des nationalen PISA-Samples (*'w_fstuwt'*) herangezogen werden. Umgekehrt kann für den Kanton Jura, wo eine Vollerhebung durchgeführt wurde, die Gewichtungsvariable des Klassensamples verwendet werden.[1]

Für die übrige französischsprachige Schweiz muss dagegen eine entsprechende Basisgewichtung erst noch konstruiert werden, welche der doppelten Auswahlchance von Jugendlichen aus der Romandie Rechnung trägt. Da es sich um zwei unabhängige Stichproben handelt, lässt sich dabei die Auswahlchance dieser Jugendlichen nach dem Additionssatz der Wahrscheinlichkeitsrechnung folgendermassen berechnen:

$$P_{Rom.i} = P_{N.i} + P_{C.i} - \left(P_{N.i} \cdot P_{C.i}\right) \qquad (2)$$

wobei:

$P_{Rom.i}$	Auswahlwahrscheinlichkeit von ProbandIn i aus der Romandie
$P_{N.i}$	Wahrscheinlichkeit von i, in die nationale Stichprobe zu gelangen
$P_{C.i}$	Wahrscheinlichkeit von i, in die Klassenstichprobe Romandie zu gelangen

Im Prinzip lässt sich die Auswahlwahrscheinlichkeit und – als deren Kehrwert – die PISA-Basisgewichtung für die Romandie problemlos berechnen, indem für $P_{N.i}$ und $P_{C.i}$ die Kehrwerte der beiden PISA-Gewichtungsvariablen für die nationale Stichprobe respektive für die Klassenstichprobe in Beziehung (2) eingesetzt werden. Die resultierende Basisgewichtung berücksichtigt damit auch bereits die in den beiden Gewichtungen enthaltenen Nonresponse-Korrekturen.

Bei der praktischen Berechnung stellt sich allerdings das Problem, dass *beide* Gewichtungsvariablen nur für jene ProbandInnen verfügbar sind, die zufälligerweise in beide Stichproben gelangt sind. Dies trifft aber bloss auf 309 der insgesamt 4943 Jugendlichen im zusammengesetzten Sample für die Romandie (ohne Jura) zu. Für 3806 Jugendliche ist demgegenüber nur die Gewichtung der Klassenstichprobe verfügbar und für 828 Jugendliche ausschliesslich jene des nationalen Samples. Für diese beiden Gruppen gilt es deshalb, die fehlende Gewichtung zu rekonstruieren.

Im Falle der *Rekonstruktion der Gewichtung für die Klassenstichprobe* ist dies aufgrund der einstufigen Auswahl und des entsprechend einfachen Gewichtungsschemas problemlos möglich. Bei der Berechnung kann von der plausiblen Annahme ausgegangen werden, dass die stratumsspezifischen Nonresponsequoten unter 828 ProbandInnen ausserhalb der Klassenstichprobe gleich ausgefallen wären wie die innerhalb der einzelnen Strata des Klassensamples empirisch beobachteten. Die Tragweite dieser Annahme ist zudem begrenzt, trägt doch

[1] Die Werte der Gewichtungsvariablen für den Kanton Jura sind wegen der Nonresponse-Korrektur ungeachtet der Vollerhebung nicht konstant und meist geringfügig grösser als 1.

die Nonresponse-Korrektur nur etwa vier Prozent zur Varianz der PISA-Gewichtungsvariablen für das Klassensample der Romandie bei.[1]

Eine *Rekonstruktion der Gewichtung fürs nationale PISA-Sample* hat sich hingegen auf Basis der verfügbaren Angaben als unmöglich erwiesen. Insbesondere wäre es für eine Nachberechnung unerlässlich, die Zahl der zur PISA-Zielpopulation zählenden SchülerInnen innerhalb der ausgewählten Schulen des Klassensamples zu kennen (Renaud, Ramseier & Zahner, 2000, 8; PISA Consortium, 2000b, 7). Diese Information ist aber nur noch für die im Rahmen des nationalen Samples faktisch ausgewählten Schulen verfügbar, nicht aber für jene, die ausschliesslich im Klassensample repräsentiert sind. Zudem ist es weder gelungen, die Information rückwirkend zu rekonstruieren noch konnte mit den verfügbaren Angaben eine approximative Rekonstruktion der nationalen Gewichtung erstellt werden.

Es bleibt damit nur noch der Ausweg einer Approximation, bei der die nicht rekonstruierbare nationale Gewichtungsvariable für die 3806 Jugendlichen der Klassenstichprobe durch den Mittelwert der nationalen Gewichtungsvariablen aller Neuntklässler aus der Romandie substituiert wird (exkl. Jura, nur Stratum 23).[2] Damit wird eine – bezogen auf das nationale Sample – für alle 3806 Jugendlichen konstante Auswahlwahrscheinlichkeit unterstellt. Obwohl diese Lösung rein theoretisch sehr unbefriedigend ist, bleiben die praktischen Auswirkungen auf die Qualität der Gewichtung gering. Für jene Neuntklässler aus der Romandie, die zur nationalen Stichprobe gehören und für welche die originale nationale Gewichtung deshalb verfügbar ist, verändert sich die resultierende Basisgewichtung jedenfalls kaum, wenn anstelle der originalen nationalen Gewichtungsvariablen der erwähnte Mittelwert in die Berechnung anhand von Beziehung (2) eingesetzt wird. Der Vergleich der Berechnungen mit und ohne Mittelwertsubstitution wird in der *folgenden Grafik* visualisiert.

Der überaus hohe Grad an Übereinstimmung ungeachtet der Vernachlässigung der an sich durchaus beträchtlichen Unterschiede in den individuellen Auswahlwahrscheinlichkeiten seitens der nationalen Stichprobe ist eine Folge davon, dass die Auswahlsätze seitens der Klassenstichprobe generell um ein mehrfaches höher liegen, weshalb die betreffende Gewichtung die anhand von Beziehung (2) ermittelte Basisgewichtung weitaus stärker beeinflusst als die nationale Gewichtungsvariable. Die leider nicht zu vermeidende Mittelwert-Substitution für die rund 3800 ProbandInnen mit fehlenden Werten bei der nationalen Gewichtungsvariablen wirkt sich also nicht nennenswert auf die Qualität der Basisgewichtung aus.

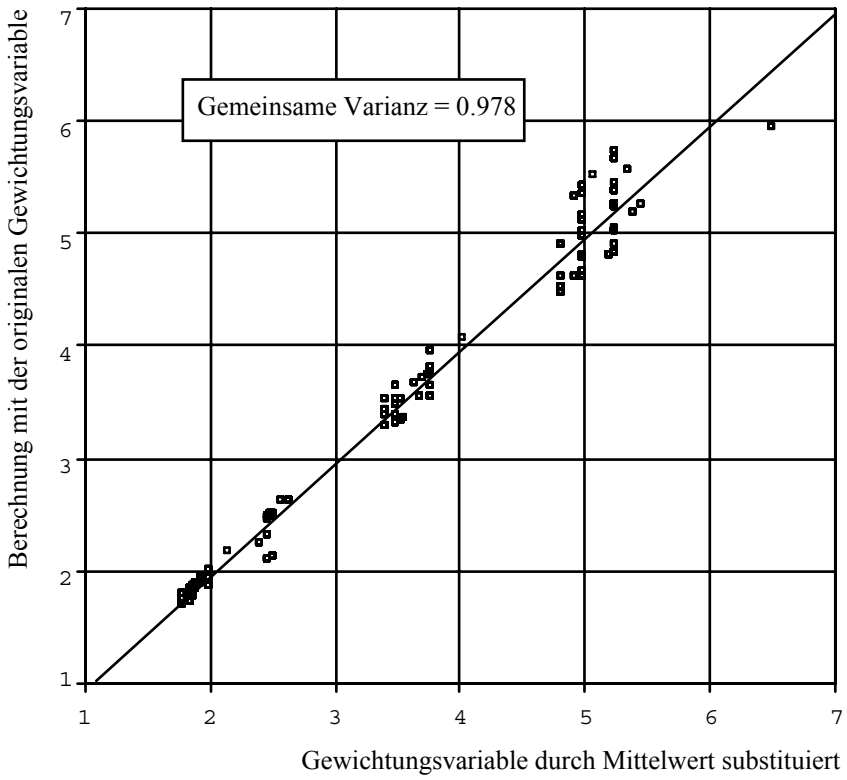
Zusammenfassend kann die Basisgewichtung also folgendermassen definiert werden:

$$G_{PISA.i} = \begin{cases} G_{Rom.i} & \text{| Stratum 22 (Kanton Jura)} \\ \left(1/G_{Rom.i} + 1/G_{Nat.i} - \left[1/G_{Rom.i} \cdot 1/G_{Nat.i} \right] \right)^{-1} & \text{| Stratum 23/61f. (übr. Romandie)} \\ G_{Nat.i} & \text{| übrige Strata (übr. Schweiz)} \end{cases} \qquad (3)$$

[1] Dies belegt eine einfaktorielle Varianzanalyse mit der Stratumsvariablen, über die die unterschiedlichen Auswahlwahrscheinlichkeiten definiert werden, als Faktor und der Gewichtungsvariablen als abhängiger Variable.

[2] Der Mittelwert der Gewichtungsvariablen beträgt 13.9.

Grafik 2: *Berechnung der Basisgewichtung für die Romandie*
Abschätzen der Auswirkungen der Mittelwertsubstitution (N=1034)



Für den *Kanton Jura*, wo eine Vollerhebung durchgeführt wurde, kann die Gewichtung der Klassenstichprobe ($G_{Rom.i}$) wie erwähnt als Basisgewichtung verwendet werden. Die Basisgewichtung für die übrige Romandie wird anhand von Beziehung (2) berechnet, während für die deutsch- und italienischsprachige Schweiz die Gewichtungsvariable des nationalen PISA-Samples ($G_{Nat.i}$) eingesetzt wird.

3.2 Teilnahmewahrscheinlichkeit bei der Adresserhebung

Wie weiter oben näher begründet, ist eine möglichst realitätsnahe Schätzung der individuellen Teilnahmewahrscheinlichkeiten an der Adressenerhebung, bei der ja der überwiegende Anteil des Nonresponse angefallen ist, für die Qualität der longitudinalen Panel-Gewichtung wesentlich (siehe Grafik 1). Es wird deshalb versucht, bei der Schätzung der Rücksendewahrscheinlichkeit mittels einer logistischen Regression (vgl. Hosmer & Lemeshow, 1989) eine möglichst umfassende Auswahl von potentiell relevanten Prädiktoren einzubeziehen. Dies ist auch angezeigt, weil Simulationsstudien darauf hinweisen, dass die Auswahl der Prädiktoren für die Qualität der Gewichtung ausschlaggebender ist, als etwa das verwendete statistische Verfahren (Rizzo, Kalton & Brick, 1994). Die PISA-ProbandInnen aus dem *Kanton Genf* bleiben von den Modellschätzungen ausgeschlossen, weil die benötigten Adressen dort mit einer (annähernden) *Vollerhebung* auf administrativem Weg erfasst worden sind. Die Modellierungen beruhen somit auf der Ausgangsstichprobe der Adressenerhebung gemäss Tabelle 1 unter Ausschluss von Genf (N = 11866).

Die Konstruktion eines erklärungskräftigen Modells des Adressblattrücklaufs orientiert sich, einer ähnlichen Philosophie wie Wießner (2003: 89) folgend, soweit möglich an Argumenten

und Befunden der Nonresponse-Forschung (Schnell, 1997; Koch & Porst, 1998), bleibt aber zwangsläufig auch teilweise induktiv. Insbesondere wird sowohl auf der Ebene der Individualmerkmale der befragten Jugendlichen als auch auf der Ebene des Schulkontextes systematisch nach Zusammenhängen zwischen den betreffenden PISA-Variablen und dem Adressblattrücklauf gesucht.[1] Sofern sich die auf diese Weise aufgedeckten Zusammenhänge theoretisch plausibel sind und sich als signifikant und statistisch robust erweisen, werden sie ins Modell integriert.

Theoretisch ist zu erwarten, dass die Bereitschaft zur Beteiligung an der Befragung und damit auch das Ausfüllen des Adressblattes von den Charakteristika der befragten SchülerInnen und ihres familiären Hintergrundes, vom schulischen Kontext und auch von situativen Faktoren abhängt, die im Rahmen der PISA-Testsessions wirksam wurden. Dabei weist die explorative Rücklaufstudie von Meyer (2000) darauf hin, dass vor allem die situativen, an die PISA-Test-Administration gebundenen Faktoren den Adressblatt-Rücklauf massiv beeinflusst haben. Zum einen liefert sie Anhaltspunkte, wonach das Streuen und Erläutern der Adresserhebungsblätter je nach EDK-Region unterschiedlich gut funktioniert hat. Dafür dürfte wohl vor allem die Qualität der Instruktion der Test-AdministratorInnen, welche die einzelnen Testsessions zu betreuen hatten, durch die regionale oder teilweise auch die kantonale Erhebungsleitung verantwortlich sein. Zum anderen ist anzunehmen, dass der Adressblatt-Rücklauf auch unterhalb der Ebene der Regionen bzw. Kantone stark von der Zusammensetzung der 'Test-Klassen' sowie vom Talent und der Motivation der einzelnen Test-AdministratorInnen abhängt, ihre ProbandInnen am Ende einer längeren Testsession nochmals zur Teilnahme zu motivieren.

Da erste explorative Modelle die Vermutung einer überragenden Bedeutung der Testadministration auf der regionalen und kantonalen Ebene vollumfänglich bestätigen, ist in einem zweiten Schritt geprüft worden, ob auch entsprechende Unterschiede zwischen den einzelnen Test-AdministratorInnen auszumachen sind. Dabei hat es sich als äusserst hinderlich erwiesen, dass die Zuordnung der ProbandInnen zu Testsessions und TestadministratorInnen im PISA-Datensatz nicht erfasst ist. Bei einer explorativen Schätzung für die EDK-Region Nordwestschweiz, wo die betreffenden Angaben relativ leicht zugänglich waren, hat sich der Einfluss der Test-AdministratorInnen indessen als der weitaus Stärkste im gesamten Modell entpuppt. Es hat sich somit aufgedrängt, die benötigten Angaben auch für die übrigen Regionen zu erschliessen, was eine aufwendige Nacherhebung in den einzelnen EDK-Regionen und beim Bundesamt für Statistik erfordert hat. Diese Nacherhebung drängte sich auch deshalb auf, weil die explorative Schätzung überzeugende Evidenz liefert, dass die starken Effekte der Testadministration auf den Rücklauf – welche für sich allein genommen als vernachlässigbar, da annähernd stichprobenneutral gelten könnten – bei Jugendlichen mit geringer Lesekompetenz noch weitaus deutlicher ausfallen. Es besteht somit eine überaus starke Interaktion zwischen den situativen, an die Testadministration gebundenen Einflüssen und einem Individualmerkmal, welches mit Blick auf die zentralen Fragestellungen von TREE in höchstem Masse relevant ist. Aus den genannten Gründen scheint es zwingend, den Einfluss der individuellen TestadministratorInnen ins Modell zu integrieren.

---

[1] Neben zahlreichen Individualmerkmalen sind im Rahmen von PISA mittels einer *Befragung der Schulen* auch vielfältige *Merkmale des schulischen Kontextes* erfasst worden (siehe PISA Consortium, 2000a). Bei der Modellentwicklung werden zunächst individuen- und kontextbezogenen PISA-Variablen, die als Prädiktoren der Beteiligung an der Adresserhebung in Betracht kommen, in ein umfassendes provisorisches Modell integriert, das dann anschliessend bereinigt wird.

Bei der dafür erforderlichen Nacherhebung hat sich allerdings gezeigt, dass die Zuordnungen von ProbandInnen und TestadministratorInnen nicht lückenlos dokumentiert sind. Nicht mehr rekonstruierbar ist die benötigte Angabe für 10,4 Prozent der Ausgangsstichprobe (N=1333), wobei sich die betroffenen ProbandInnen überwiegend in den EDK-Regionen Zürich (N=461), übrige Deutschschweiz (N=481) und italienischsprachige Schweiz (N=379) konzentrieren.[1] Da in den meisten Schulen bloss eine Testsession durchgeführt worden ist, kann in diesen Fällen die Zuordnung der Befragten zu den TestadministratorInnen näherungsweise durch den Identifikator der Schulen im Datensatz (Variable 'schoolid') substituiert werden.[2] Des Weiteren gilt es zu berücksichtigen, dass eine Reihe von Testsessions von Teams von zwei oder mehr Personen betreut worden ist. Da nicht bekannt ist, wie sich dies auf situativen Einflüsse im Rahmen der Testsessions ausgewirkt hat, werden die Teams, die 3,5 Prozent der ProbandInnen betreut haben (N=454), bei der Modellschätzung wie die AdministratorInnen als je eigene Kategorien – als 'synthetische' Personen sozusagen – spezifiziert.

Bei der *Schätzung der logistischen Regression* unter Einschluss des rekonstruierten Identifikators für die TestadministratorInnen stellt sich allerdings das Problem, dass überaus zahlreiche Parameter zu schätzen sind, was unweigerlich zu relativ ungenauen Schätzungen führt. Die Identifikator-Variable hat nämlich 163 Kategorien, die zwischen lediglich 6 und immerhin 246 Befragten umfassen – durchschnittlich rund 73 – und für die je zwei Effektparameter zu schätzen sind, sofern die starke Interaktion zwischen dem Einfluss der AdministratorInnen und der individuellen Lesekompetenz ins Modell integriert wird. Hinzu kommen die Effektparameter der übrigen Individual- und Kontextvariablen im Modell, so dass das provisorische Modell rund 380 zu schätzende Parameter aufweist. Da die Präzision der Effektschätzungen aber die Genauigkeit der geschätzten Antwortwahrscheinlichkeiten und damit letztlich auch der longitudinalen Gewichtung massgeblich beeinflusst, drängt sich eine stärkere Aggregation solcher TestadministratorInnen auf, die im Durchschnitt wie auch für ProbandInnen mit vergleichbarer Lesekompetenz eine gleich gute Beteiligung an der Adressenerhebung erreicht haben.

Für die Zusammenfassung von AdministratorInnen mit vergleichbarer 'Performance', die möglichst keine substanzielle Verschlechterung der Modellanpassung nach sich ziehen sollte, wird eine Clusteranalyse durchgeführt (Everitt, 1993).[3] Dabei bilden die 163 TestadministratorInnen die Analyse-Einheiten und die Effektparameter (B) des provisorischen Regressionsmodells, welche den individuellen Einfluss der einzelnen AdministratorInnen sowie die Stärke der Interaktion zwischen diesem Einfluss und Lesekompetenz abbilden, die verwendeten Variablen.[4] In *Grafik 3* sind die beiden Variablen gegeneinander aufgetragen, wobei sich sehr deutlich ein theoretisch einleuchtendes Muster zeigt: Wo es den TestadministratorInnen

---

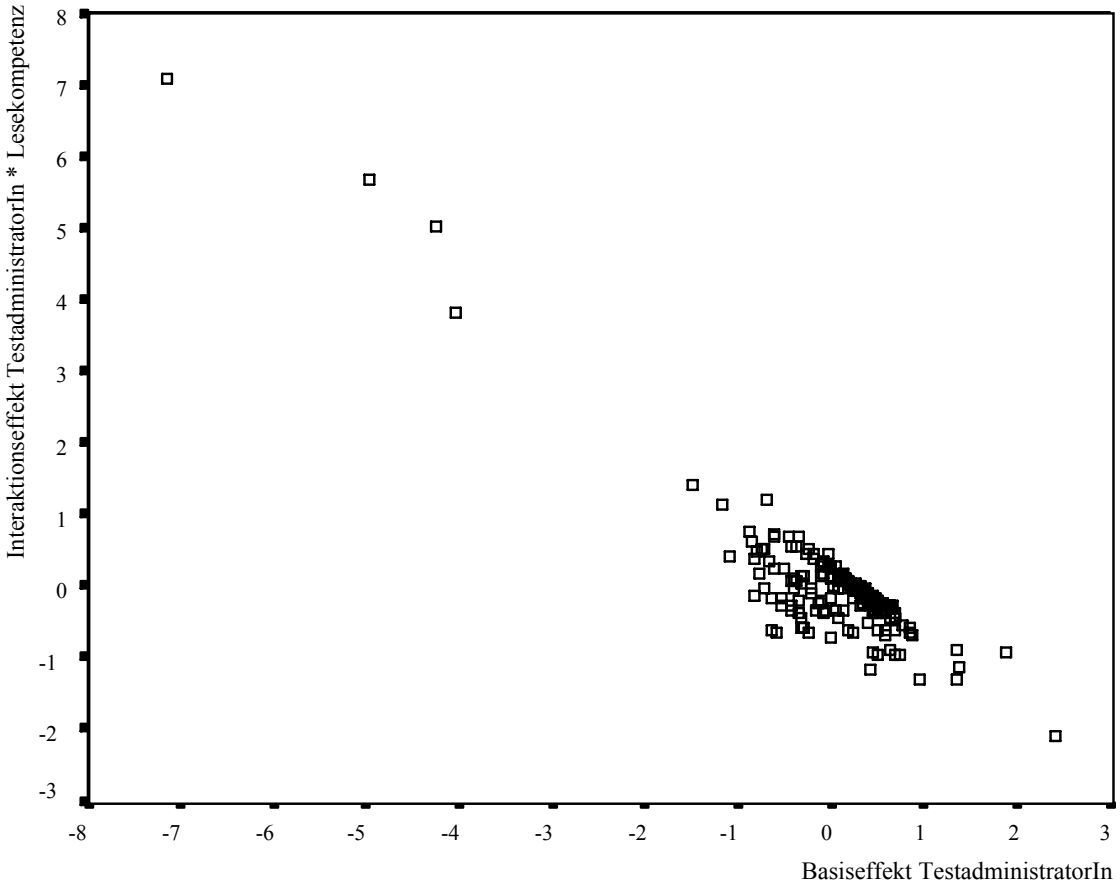
1 Innerhalb der Erhebungsregion 'Deutschschweiz ohne Zürich' bildet der Kanton Sankt Gallen mit 323 fehlenden Zuordnungen den Schwerpunkt.

2 Im Sinne eine Fehlerminimierung ist davon auszugehen, dass damit zwar SchülerInnen, die von derselben TestadministratorIn betreut wurden, verschiedenen Kategorien zugeteilt werden, was aber weniger gravierend scheint, als sie in einer heterogenen, von mehreren TestadministratorInnen betreuten Sammelkategorie zu belassen.

3 Die Veränderung der Modellanpassung aufgrund von unterschiedlichen Aggregationsverfahren und -varianten wird anhand von Likelihood-Ratio-Tests ermittelt (vgl. Hosmer & Lemeshow, 1989).

4 Da beide Effekte auf Intervallniveau gemessen sind, wird die Clusteranalyse auf der Basis von euklidischen Distanzen nach der Ward-Methode durchgeführt. Die beiden Variablen sind vorgängig z-standardisiert worden, was ihnen bei der Clusteranalyse einen gleich starken Einfluss auf den Agglomerationsprozess verleiht.

Grafik 3: *Zusammenhang zwischen dem Einfluss der Qualität der Testadministration (Basiseffekt) und dem Einfluss der Lesekompetenz (Interaktionseffekt)*



insgesamt schlecht gelingt, die 'Test-Klassen' zur Teilnahme zu motivieren – was sich an einem negativen Basiseffekt ablesen lässt –, findet sich zugleich durchwegs ein Interaktionseffekt, der einen starken, positiven Einfluss der Lesekompetenz auf die Beteiligung an der Adressenerhebung anzeigt. Eine unzureichende Motivierung durch die TestadministratorInnen drückt somit den Rücklauf in erster Linie bei Jugendlichen mit geringer Lesekompetenz, während in Testklassen mit durchschnittlich sehr guter Beteiligung keinerlei Zusammenhang zwischen Lesekompetenz und Beteiligung auszumachen ist.[1] Die dargestellten individuellen Interaktions- und Basiseffekte sind übrigens – ungeachtet der schwachen Besetzung vieler Kategorien – mehrheitlich signifikant. Bezogen auf das provisorische Gesamtmodell haben Testadministration und ihre Interaktion mit der Lesekompetenz die insgesamt stärksten Einflüsse.[2]

Auf der Basis der Clusteranalyse ist es möglich, die TestadministratorInnen zu nur noch 18 ausreichend besetzten Kategorien zu gruppieren, ohne dass dadurch eine substanzielle Verschlechterung der Modellanpassung in Kauf zu nehmen wäre.[3] Nach der Zusammenfassung

---

1 Die Effektschätzungen sind als Abweichungen von den Werten jenes Testadministrators zu verstehen, der die Referenzkategorie bildet und für den der Zusammenhang zwischen Lesekompetenz und Rücklauf sehr nahe bei Null liegt und nicht signifikant ist.

2 Dies gemessen am Beitrag zum χ^2-Wert des Gesamtmodells.

3 Ausgehend von der Lösung mit 40 Clustern sind noch weitergehende manuelle Zusammenfassungen vorgenommen worden, wobei jeweils mit einem Likelihood-Ratio-Test sichergestellt wird, dass sich die Modellanpassung dabei nicht verschlechtert. Das bevorzugte Modell mit letztlich noch 18 Testadministrations-Kategorien ($N \geq 67$) hat einen χ^2-Wert von 3321 bei 63 Freiheitsgraden, während das provisorische Aus-

resultiert das in *Tabelle 3* wiedergegebene, nun definitive Modell zur Schätzung der individuellen Teilnahmewahrscheinlichkeiten an der Adresserhebung ($A_{ADR,i}$).

In der definitiven Modellschätzung sind angesichts der sehr grossen Stichprobe bloss noch Variablen berücksichtigt, deren Einfluss bei einer *Irrtumswahrscheinlichkeit* (α) von einem Prozent gesichert ist. Dies schliesst bei kategorialen Prädiktoren nicht immer aus, dass die Effektkoeffizienten einzelner Kategorien nicht dieses Signifikanzniveau erreichen. Wo dies aufgrund der provisorischen Effektschätzungen angezeigt scheint und, gemessen am Likelihood-Ratio-Test, ohne signifikante Verschlechterung des Modells möglich ist, werden die einzelnen Kategorien der Prädiktoren auf Nominalniveau zudem stärker zusammengefasst. Dies verbessert die Präzision der Effektschätzungen und damit der Gewichtung. Alle aufgeführten Effekte sind nicht nur statistisch gesichert, sondern auch auf *Robustheit* geprüft.[1] Die verwendeten *Rekodierungen* sind in Tabelle 3 unmittelbar neben den Variablen-Etiketten in Klammern oder dann im Tabellenfuss wiedergegeben.

Wendet man sich zuerst den *Kennwerten für das Gesamtmodell* zuunterst in *Tabelle 3* (Fortsetzung der Tabelle auf Seite 16) zu, so zeichnet sich dieses durch eine sehr gute Anpassung an die Daten (Hosmer-Lemeshow-Test) und einen alles in allem recht engen Zusammenhang zwischen Nonresponse und Prädiktoren aus (Likelihood-Ratio- oder McFadden-Pseudo-R^2).[2] Ein Blick auf die Bedeutung der einzelnen Variablenblöcke für die Modellanpassung, der sich anhand der Aufteilung der Likelihood-Beiträge abschätzen lässt, zeigt im weiteren, dass der weitaus grösste Teil der systematischen Beteiligungsunterschiede auf die Einflüsse der Testadministration ($\approx$ 1100 χ^2-Punkte) zurückgeht. Daneben finden sich ebenfalls substanzielle, aber weniger bedeutende Einflüsse des regionalen und schulischen Kontextes ($\approx$ 250 χ^2-Punkte) und von individuellen und familiären Merkmalen ($\approx$ 320 χ^2-Punkte) .

Die dominierenden *Einflüsse von Testadministration und Sampling* sind wie erwartet auf zwei Ebenen anzusiedeln. Zum einen bestehen bedeutende Unterschiede zwischen den EDK-Regionen, die wohl hauptsächlich dem Einfluss der regionalen PISA-Erhebungsleitung zuzuschreiben sind. Dieser ist in Zürich und vor allem in der Romandie (exkl. Genf) äusserst ungünstig. Zum anderen findet sich auch im definitiven Modell der bereits beschriebene, überaus starke *Einfluss der einzelnen TestadministratorInnen*. Dieser ist am ungünstigsten für Cluster 1 und 2, wo der Adressblatt-Rücklauf weitaus am tiefsten liegt und zugleich weitaus am stärksten von der individuellen Lesekompetenz gemäss PISA (Variable *'wleread'*) abhängt, wie die Interaktionseffekte belegen. Am anderen Pol der 18 Testadministrations-Kategorien finden sich dagegen eine Reihe von Clustern mit einem überaus hohen Rücklauf, der zudem auch kaum von der Lesekompetenz beeinflusst ist. In einigen Clustern (insb. Nr. 17)

---

gangsmodell 3523 χ^2-Punkte bei 380 Freiheitsgraden aufweist. Die resultierende Verschlechterung der Modellanpassung ist somit gemäss Likelihood-Ratio-Test nicht signifikant (Hosmer & Lemeshow, 1989, 105), und das resultierende Modell erweist sich auch gemessen am BIC (Hagenaars, 1990: 67) als optimal.

[1] Dafür wurden separate Modellschätzungen unter Ausschluss der Fälle mit den höchsten Cook- respektive DfBeta-Werten durchgeführt.

[2] Nach Menard (2002: 24-26) ist das Likelihood-Ratio Pseudo-R^2 − auch als McFadden-R^2 bezeichnet − aus mehreren Gründen die bestgeeignete Kennzahl, um in einer logistischen Regression die Gesamtstärke der Assoziation zwischen Prädiktoren und abhängiger Variable zu bestimmen (zur Berechnung: ibd., Seite 24).

Tabelle 3: *Schätzung Teilnahmewahrscheinlichkeit Adressenerhebung*

Logistische Regression (N=11866)		B	S.E.	Wald	df	Sig.	Exp(B)
Einflüsse von Testadministration & Sampling							
Erhebungsregion [K]				464.3	3	.000	
Zürich	(Reg_5=2)	– 1.281	.082	245.0	1	.000	.278
Romandie (ohne Genf)	(Reg_5=3)	– 2.625	.214	151.0	1	.000	.072
Italienischsprachige Schweiz	(Reg_5=5)	.659	.105	39.7	1	.000	1.932
TestadminstratorInnen gruppiert [K]				270.8	17	.000	
Teilnahme minimal	(Cluster 1)	– 10.788	2.943	13.4	1	.000	.000
…	(Cluster 2)	– 8.829	3.704	5.7	1	.017	.000
…	(Cluster 3)	– 4.931	.561	77.2	1	.000	.007
…	(Cluster 4)	– 4.062	.975	17.4	1	.000	.017
…	(Cluster 5)	– 3.467	.856	16.4	1	.000	.031
…	(Cluster 6)	– 3.440	.666	26.7	1	.000	.032
…	(Cluster 7)	– 2.210	.469	22.2	1	.000	.110
…	(Cluster 8)	– 1.575	.399	15.6	1	.000	.207
…	(Cluster 9)	– 1.600	.297	28.9	1	.000	.202
…	(Cluster 10)	– 1.608	.717	5.0	1	.025	.200
…	(Cluster 11)	– .376	.376	1.0	1	.318	.687
…	(Cluster 13)	.759	.555	1.9	1	.172	2.137
…	(Cluster 14)	.982	.658	2.2	1	.136	2.669
…	(Cluster 15)	1.417	.418	11.5	1	.001	4.124
…	(Cluster 16)	4.327	1.417	9.3	1	.002	75.689
…	(Cluster 17)	5.056	1.396	13.1	1	.000	157.038
Teilnahme maximal	(Cluster 18)	7.468	2.000	13.9	1	.000	1751.468
Interaktion: Testadministration * Lesekompetenz [1)]				318.2	17	.000	
Interaktion stark positiv	(Cluster 1)	.019	.006	11.8	1	.001	1.019
…	(Cluster 2)	.025	.008	10.2	1	.001	1.025
…	(Cluster 3)	.011	.001	98.1	1	.000	1.011
…	(Cluster 4)	.007	.002	12.9	1	.000	1.007
…	(Cluster 5)	.009	.002	27.5	1	.000	1.009
…	(Cluster 6)	.004	.001	12.4	1	.000	1.004
…	(Cluster 7)	.005	.001	30.9	1	.000	1.005
…	(Cluster 8)	.005	.001	44.2	1	.000	1.005
…	(Cluster 9)	.003	.001	28.7	1	.000	1.003
…	(Cluster 10)	.007	.001	24.3	1	.000	1.007
…	(Cluster 11)	.002	.001	7.8	1	.005	1.002
…	(Cluster 13)	– .003	.001	6.6	1	.010	.997
…	(Cluster 14)	.001	.001	.4	1	.522	1.001
…	(Cluster 15)	– .001	.001	2.0	1	.163	.999
…	(Cluster 16)	– .005	.003	3.9	1	.049	.995
…	(Cluster 17)	– .010	.003	12.1	1	.001	.990
Interaktion schwach negativ	(Cluster 18)	– .006	.004	2.4	1	.121	.994
Teilstichprobe / Sampling [K]				34.5	2	.000	
Klassenstichprobe Romandie	(sourcnat=2)	.746	.206	13.1	1	.000	2.108
Nat. Sample in PSU Klassensample	(sourcnat=1, scoolsmp=3)	1.225	.227	29.2	1	.000	3.403
Kontextmerkmale: Region, Schulhaus, Lehrer							
Kanton Luzern [D]	(canton = 3)	– .407	.135	9.1	1	.003	.666
Kanton Thurgau [D]	(canton = 20)	– 1.268	.157	64.9	1	.000	.281
Kanton Wallis [D]	(canton = 23)	– .727	.118	38.2	1	.000	.483
Schule: Anzahl Schüler [2)]		– .178	.038	22.2	1	.000	.837
Schultyp: Privatschule / Angabe fehlt [D]	(sc03q01= 2, 7, 8, 9, .)	– .683	.089	59.5	1	.000	.505
Schule: Häufigkeit standardisierter Tests [K]				40.3	2	.000	
zirka 2 pro Jahr	(sc16q01= 3)	– .658	.112	34.4	1	.000	.518
3 oder mehr pro Jahr	(sc16q01= 4, 5)	.204	.095	4.6	1	.032	1.226
Einsatz des Lehrers [3)]		.099	.030	11.0	1	.001	1.104

Tabelle 3: *Fortsetzung*		B	S.E.	Wald	df	Sig.	Exp(B)
Individualmerkmale							
Schultyp: Integriert oder Gymnasium [D]	(typ = 4, 7)	– .370	.085	19.0	1	.000	.691
Nicht im 9. Schuljahr [D]	(source915 = 2)	.398	.084	22.3	1	.000	1.488
Note in Testsprache: Angabe fehlt [D]	(st41q04 = 7, 8, 9, .)	– .585	.089	43.6	1	.000	.557
Bekommt Nachhilfeunterricht [D]	(st23q01 ≥ 1)	.140	.050	7.8	1	.005	1.151
Lesestunden pro Tag [K]				16.0	2	.000	
weniger als eine Stunde	(st34q01 = 2, 3)	.162	.051	10.0	1	.002	1.176
eine Stunde oder mehr	(st34q01 = 4, 5)	.408	.136	9.0	1	.003	1.504
Nutzt gerne Buchhandel / Bibliotheken [K]				40.6	2	.000	
eher nein | eher ja	(st35q07 = 2, 3)	.198	.053	14.1	1	.000	1.219
ja	(st35q07 = 4)	.495	.078	40.5	1	.000	1.641
Wäre gerne irgendwo der Beste [K]				38.1	2	.000	
stimmt eher	(cc02q16 = 3)	.146	.052	7.9	1	.005	1.157
stimmt genau	(cc02q16 = 4)	.327	.053	38.1	1	.000	1.386
Projektarbeit: Ideen aller sammeln [D]	(cc02q22 = 4)	.156	.047	11.0	1	.001	1.169
Familie besitzt Gedichtbände [D]	(st21q10 = 1)	.184	.045	16.6	1	.000	1.202
Anzahl ältere Geschwister [K]				17.4	2	.000	
eins bis zwei	(st05q01 = 2, 3)	– .137	.044	9.8	1	.002	.872
drei oder mehr	(st05q01 = 4, 5)	– .344	.101	11.5	1	.001	.709
Geburtsland: nicht die Schweiz [D]	(st16q03 = 2)	– .164	.050	10.8	1	.001	.849
Geschlecht: Weiblich [D]	(st03q01 = 1)	.320	.045	50.7	1	.000	1.377
Regressionskonstante		.771	.250	9.5	1	.002	2.162
Modellkennwerte							
– 2 Log-Likelihood		13112					
Model-χ^2 (df) P		3321	(63)	.00000			
Likelihood-Ratio Pseudo R^2 (McFadden)		.202					
Model-Fit: Hosmer-Lemeshow Test: [χ^2, (df), P]		6.0	(8)	.643			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). K: Kategoriale Variable: Die nicht explizit ausgewiesenen Kategorien (teils inkl. fehlende Werte) bilden die Referenzkategorie. 1) Die Lesekompetenz (PISA-Variable *'wlreread'*) ist zentriert, was dazu beiträgt, die Koeffizientenschätzungen zu stabilisieren (Jaccard, 2001); 19 fehlende Angaben sind durch den Mittelwert substituiert. 2) Natürlicher Logarithmus aus der Summe der Gesamtzahl an Schülern (*'sc02q01'*) und Schülerinnen (*'sc02q02'*); 841 fehlende Angaben sind durch den Mittelwert substituiert. 3) Mittelwert aus Item st26q05 bis st26q10 (max. 2 Item missings erlaubt, 15 fehlende Skalenwerte durch Mittelwert substituiert).

haben sich die leseschwachen ProbandInnen sogar leicht häufiger beteiligt.[1] Das *Muster* der beobachteten Effekte entspricht damit weiterhin demjenigen des provisorischen Modells mit separaten Effektschätzungen für einzelnen TestadministatorInnen, so wie es in Grafik 3 dargestellt ist.

Im Weiteren hat auch die vom Sampling abhängige *Art der Testsession* einen bedeutenden Einfluss auf die Antwortwahrscheinlichkeit. Erstens haben die SchülerInnen des Klassensamples Romandie, die im gewohnten Rahmen ihrer Schulklasse befragt wurden, weitaus öfter geantwortet, als jene des nationalen Samples, bei dem SchülerInnen aus verschiedenen Klassen zu Testsessions zusammengezogen wurden. Zweitens haben jene SchülerInnen am besten geantwortet, die ausschliesslich zum nationalen Sample zählen, in deren Schulhaus aber auch ganze Klassen befragt wurden. Diese SchülerInnen sind in der Regel nicht einzeln befragt

[1] Für die Referenzkategorie der TestadministratorInnen-Gruppierung (Cluster 12) – wie auch für die meisten anderen höherrangigen Cluster – liegt kein nennenswerter Einfluss der Lesekompetenz vor. Der Indikator für die Lesekompetenz ist übrigens aus dem definitiven Modell ausgeschlossen worden, weil der Effektkoeffizient vernachlässigbar klein und dementsprechend auch nicht signifikant ist.

worden, sondern zusammen mit einer für sie fremden Klasse des Klassensamples. Die im Rahmen der Klassenbefragungen offensichtlich sehr günstigen situativen Faktoren haben dabei bei den zugeteilten SchülerInnen offenbar noch stärker gewirkt, möglicherweise weil diese sich innerhalb der 'fremden' Klasse konformer verhalten.

Neben diesen unmittelbar an Testadministration und Sampling geknüpften Faktoren haben *weitere* Kontextbedingungen die Beteiligung an der Adress-Erhebung beeinflusst. Dies trifft zunächst auf den *räumlichen Kontext* zu. Für drei *Kantone* findet sich ein deutlich geringerer Adressen-Rücklauf, allen voran für den Thurgau, aber auch fürs Wallis und Luzern. Teilweise kann dies ebenfalls auf Testadministrationseinflüsse zurückgeführt werden, bilden doch die Kantone eine Art mittlere Ebene zwischen der regionalen Erhebungsleitung und den einzelnen TestadministratorInnen. Zumindest im Falle des Kantons Wallis, für den sich bei einer separaten Berechnung für Ober- und Unterwallis praktisch identische Effekte finden, obschon die beiden Kantonsteile zu zwei verschiedenen Testregionen gehören, scheidet diese Erklärung aber aus. Denkbar ist auch, dass sich die kantonalen Unterschiede in der Einstellung von Lehrkräften und/oder ProbandInnen gegenüber Umfragen und wissenschaftlichen Studien auf die Beteiligung ausgewirkt haben.

Des Weiteren spielen auch *Kontextbedingungen* eine Rolle, die an *Schulhaus und Lehrperson* geknüpft sind. So nimmt erstens die Beteiligung mit der *Grösse des Schulhauses* ab. Zweitens haben sich SchülerInnen in *Privatschulen* sowie in Schulen unbekannten Typs weniger beteiligt als andere. Drittens ergibt sich für Schulen, die gelegentlich *standardisierte Tests* durchführen, eine deutlich tiefere Beteiligung als für Schulen, die selten oder nie, oder dann aber sehr häufig solche Tests durchführen. Schulen mit der grössten Testhäufigkeit heben sich dabei ihrerseits noch leicht positiv von denen mit der geringsten ab. Die Unterschiede sind wohl in Ähnlichkeiten zwischen der PISA-Befragung, die unter anderem ja auch standardisierte Leistungstests beinhaltet, und den sonst von den Schulen eingesetzten Tests begründet. Dabei scheint plausibel, dass die für das Rücksenden des Adressblattes entscheidende Bereitschaft, sich freiwillig an solchen Tests zu beteiligen, dort besonders hoch ist, wo solche Test entweder einen gewissen Neuigkeitswert aufweisen oder wo sie im Gegenteil bereits Routine sind und entsprechend schnell ausgefüllt werden können. Zu guter Letzt spielt auch das *Engagement der Lehrperson* eine Rolle: SchülerInnen, die ihre Lehrer im Schulalltag als interessiert und unterstützend erleben, stellen ihre Adresse häufiger zur Verfügung als andere.

Betrachtet man schliesslich die Ergebnisse zu den *Individual- und Familienmerkmalen,* so findet sich eine ganze Reihe von substanziellen und meist auch hochsignifikanten Effekten. SchülerInnen in *Gymnasien* sowie solche in *integrierten Klassen* der Sekundarstufe I beteiligen sich etwas weniger häufig als andere, ebenso die SchülerInnen der *9. Klasse* im Vergleich zu jenen 15jährigen im Sample, die sich in tieferen oder höheren Klassen befinden. Jugendliche, die im Rahmen der PISA-Befragung ihre *Note in der Testsprache* nicht angeben, beteiligen sich dementsprechend auch deutlich seltener an der Adressenerhebung. Befragte, die in den Jahren vor der Befragung *Nachhilfeunterricht* erhalten haben, beteiligen sich überdurchschnittlich an der Adressenerhebung. Ebenso nehmen *bibliophile Jugendliche* erheblich häufiger teil, was sich sowohl an der Zahl der täglichen *Lesestunden* wie auch an der *Nutzung von Bibliotheken und Buchhandel* ablesen lässt. ProbandInnen, die gemäss ihrer Selbsteinschätzung eher *ehrgeizig* ('wäre irgendwo gerne der Beste'), sowie *kooperativ* und *konsensorientiert* ('Ideen aller sammeln') sind, stellen ihre Adresse öfter zur Verfügung. Auch Jugendliche, deren Familie *Gedichtbände* besitzt, beteiligen sich spürbar häufiger, was im Bourdieuschen Sinne als Effekt eines spezifischen *elterlichen Bildungsmilieus* mit viel Kulturkapital gedeutet werden kann. Weiter spielt zwar die Geschwisterzahl keine Rolle, wohl aber die die

Geschwisterfolge: Die Beteiligung sinkt mit der Zahl der *älteren* Geschwister recht deutlich, was sich – in Einklang mit der Literatur – mit einem daran geknüpften, einer geringeren Konformität förderlichen Sozialisationseffekt erklären lässt. Schliesslich liegt die Beteiligung junger *Frauen* erheblich höher als jene der Männer, während jene der ProbandInnen *ausländischer Herkunft* erfreulicherweise nur leicht tiefer ausfällt.

Insgesamt ist das Ergebnis mit Blick auf eine leistungsfähige Gewichtung erfreulich. *Zum einen* ist es gelungen, ein gut angepasstes Modell zu konstruieren. Es dürfte eine vergleichsweise effektive Korrektur von Nonresponse-Verzerrungen gewährleisten, sind doch potenzielle Einflüsse der zahlreichen von PISA erhobenen Merkmale auf den Rücklauf relativ umfassend geprüft und gegebenenfalls ins Modell integriert worden. Die verfügbaren PISA-Merkmale decken zudem das thematische Spektrum der auch für den TREE-Längsschnitt relevanten Charakteristika bereits recht gut ab. *Zum anderen* lässt sich mit Blick auf die überwiegenden Einflüsse der Testadministration festhalten, dass die Unterschiede in den individuellen Beteiligungswahrscheinlichkeiten hauptsächlich auf situative Faktoren zurückgehen, deren Wirkung bezüglich der TREE-relevanten Individual-, Familien- und Kontextmerkmale als annähernd stichprobenneutral einzuschätzen ist. Von der Interaktion mit der Lesekompetenz einmal abgesehen, beeinflussen die dominanten Testadministrationseffekte ja unterschiedslos *alle* ProbandInnen.

Die geschätzten Teilnahmewahrscheinlichkeiten können nun anhand der Effektkoeffizienten und der individuellen Variablenwerte problemlos gemäss folgendem Ausdruck berechnet (vgl. Menard, 2002: 13) und dann in Beziehung (1) eingesetzt werden:

$$A_i = \frac{e^{\left[B_0 + \sum_{j=1}^{J} B_j \cdot X_{ji}\right]}}{1 + e^{\left[B_0 + \sum_{j=1}^{J} B_j \cdot X_{ji}\right]}} \qquad (4)$$

wobei:

A_i	Geschätzte Wahrscheinlichkeit, dass ProbandIn i teilnimmt
B_0	Regressionskonstante (gemäss Tabelle)
B_j	Effektkoeffizient von Variable j (gemäss Tabelle)
X_{ji}	Ausprägung von Variable j für ProbandIn i

Im Kanton Genf sind die Adressen praktisch vollzählig auf administrativem Weg erhoben worden, weshalb dieser Kanton von der Schätzung des Regressionsmodells ausgeschlossen bleibt. Anstelle der Teilnahmewahrscheinlichkeit gemäss Beziehung 4 wird deshalb für die Jugendlichen aus Genf eine Konstante in Beziehung 1 eingesetzt, die wegen vereinzelter Ausfälle minim von 1 abweicht.[1]

[1] Die Konstante beträgt 0.987 (916 gültige Adressen ÷ 928 ProbandInnen) .

3.3 Teilnahmewahrscheinlichkeit Welle 1

Das Vorgehen bei der Schätzung der individuellen Wahrscheinlichkeiten einer Beteiligung an der ersten Befragungswelle entspricht genau demjenigen beim Adressblatt. Zuerst wird geprüft, welche Variablen des Adressblattmodells auch für die Teilnahme an Welle 1 relevant sind. Anschliessend wird unter den PISA-Merkmalen systematisch nach weiteren möglichen Prädiktoren gesucht.

Das zu Grunde liegende Sample ist die *bereinigte* TREE-Ausgangsstichprobe (N = 6343, vgl. Tabelle 1), umfasst also jene Jugendlichen, die zur Untersuchungspopulation gehören und deren Adresse bekannt ist. Als Teilnehmende gelten dabei jene 5767 Samplemitglieder, die entweder den schriftlichen *Fragebogen* ausgefüllt haben, oder mit denen andernfalls mindestens ein *Telefoninterview* geführt worden ist. Die Ergebnisse sind in *Tabelle 4* wiedergegeben.

Betrachtet man zunächst die *Kennwerte fürs Gesamtmodell*, so erweist sich dieses als ähnlich gut an die Daten angepasst wie das Modell für den Adressblatt-Rücklauf (vgl. Hosmer-Lemeshow-Test), es weist aber ein erheblich geringeres Pseudo-R^2 auf (McFadden), was als (erfreulicher) Hinweis auf eine unsystematischere Verteilung des Nonresponse gewertet werden kann. Gemessen an den Likelihood-Beiträgen fällt das relative Gewicht der drei Variablenblöcke nun allerdings erwartungsgemäss radikal anders aus: Die Beiträge der Testadministration (74 χ^2-Punkte) und der Kontextbedingungen (≈ 26 χ^2-Punkte) zur Modellanpassung sind verglichen mit dem der Individualmerkmalen (≈ 312 χ^2-Punkte) nunmehr relativ unbedeutend. Dies bestätigt eindrücklich die einleitend formulierte Vermutung, wonach ganz andere Faktoren über das Ausfüllen des Adressblattes bestimmen als über die Teilnahme an der ersten Befragungswelle. Die separate Modellierung beider Ausfallprozesse trägt also wie vermutet erheblich zur Präzision der Längsschnittgewichtung bei.

Die *Testadministration* bzw. das Sampling hat wie erwähnt erwartungsgemäss einen erheblich geringeren Einfluss auf die Beteiligung an Welle 1. Insbesondere ist nun nicht mehr relevant, an welcher PISA-Testsession die Jugendlichen teilgenommen haben. Auffallend ist, dass die Romandie (exkl. Genf) eine überdurchschnittliche Beteiligung aufweist, was den stark negativen Effekt bei der Adressen-Erhebung tendenziell kompensiert. Jene Befragten aus der Romandie, welche ihre Adresse ungeachtet der dort besonders ungünstigen situativen Einflüsse der Testadministration zur Verfügung gestellt haben, bilden offensichtlich eine überdurchschnittlich interessierte und motivierte Gruppe. Ein spiegelbildlicher Mechanismus dürfte in Genf spielen, wo eine deutlich unterdurchschnittliche Beteiligung resultiert: Da die Adressen dort annähernd vollzählig auf administrativen Weg erhoben worden sind, haben nicht teilnahmebereite ProbandInnen vor der ersten Befragungswelle keine Gelegenheit, die Mitarbeit zu verweigern. Das Ergebnis der Adressenerhebung findet somit für die beiden Regionen mit der höchsten bzw. tiefsten Erfolgsquote einen gegenläufigen, kompensatorisch wirkenden ‘Nachhall’ in der Beteiligung an der ersten Befragungswelle.

Innerhalb der Deutschschweiz und innerhalb der Romandie bestehen zudem substanzielle *kantonale Beteiligungsdifferenzen*. Im definitiven Modell sind dabei Kantone mit – ceteris paribus – etwa gleicher Beteiligung zusammengefasst. Im deutschsprachigen Landesteil fallen dabei in erster Linie Glarus, Luzern, Nidwalden, beide Basel, der Thurgau sowie das Oberwallis durch eine erheblich geringere Beteiligung auf. Innerhalb der Romandie (exkl. Genf) fallen die Waadt, Neuchâtel sowie das Unterwallis erheblich ab. Worauf diese Differenzen im Einzelnen zurückgehen, lässt sich nicht ohne weiteres erkennen. Die Kantone mit einer geringeren Beteiligung unterscheiden sich punkto Wirtschaftsstruktur und Urbani-

Tabelle 4: *Schätzung Teilnahmewahrscheinlichkeit Befragungswelle 1*

Logistische Regression (N=6343)		B	S.E.	Wald	df	Sig.	Exp(B)
Testadministration							
Erhebungsregion [K]				73.7	2	.000	
Romandie ohne Genf	(Reg_5=3)	.620	.193	10.4	1	.001	1.859
Genf	(Reg_5=4)	– .755	.104	52.6	1	.000	.470
Räumlicher Kontext							
Kanton Neuchâtel & Unterwallis [D]	(canton = 24, 23: Reg_5=3)	– .707	.220	10.3	1	.001	.493
Kanton Waadt [D]	(canton = 22)	– .859	.240	12.9	1	.000	.423
Kantone LU, NW, GL, BS, BL, TG & Oberwallis [D]	(canton = 3, 7, 9, 12, 13, 20, 23: Reg_5=1)	– .456	.139	10.7	1	.001	.634
Individualmerkmale							
Nicht im 9. Schuljahr [D]	(source915 = 2)	– .956	.283	11.4	1	.001	.384
Lesekompetenz [1)]	(wleread)	.003	.000	45.4	1	.000	1.003
Mathematiknote: ungenügend, Missing [D]	(st41q05 = 3, 7, 8, 9, .)	– .350	.098	12.7	1	.000	.705
Pläne für nächstes Jahr [K]				22.4	2	.000	
Zwischenjahr, 8. / 9. Klasse	(st42q01 = 1, 2)	.399	.128	9.7	1	.002	1.490
Job, unbestimmt	(st42q01 = 13, 97)	– 1.100	.324	11.6	1	.001	.333
Schulbesuch: Negative Einst. (Rating) [2)]	(st22q01)	– .133	.038	12.4	1	.000	.876
Hausaufgaben rechtzeitig fertig: Nie [D]	(st32q01 = 1)	– .541	.169	10.3	1	.001	.582
Rock-, Pop-, Technoveranst.: ≥ 4 / Jahr [D]	(st18q03 = 4)	– .487	.120	16.4	1	.000	.614
Zuhause: Anz. Mobiltelefones (Rating) [3)]	(st22q01)	– .160	.039	17.3	1	.000	.852
Zuhause: Anzahl Taschenrechner (≥3) [D]	(st22q03 = 4)	.327	.093	12.3	1	.000	1.387
Lebt nicht mit Mutter [D]	(st04q01 = 2)	– .547	.182	9.1	1	.003	.579
Lebt nicht mit Vater [D]	(st04q03 = 2)	– .262	.102	6.6	1	.010	.770
Geburtsland ausserhalb Mitteleuropa [D]	(st16n01 ≥ 5)	– .404	.105	14.9	1	.000	.667
Geschlecht: Weiblich [D] (bereinigt)	(st03q01 = 1)	.346	.081	18.4	1	.000	1.413
Regressionskonstante		1.093	.288	14.4	1	.000	2.985
Modellkennwerte							
– 2 Log-Likelihood		4384					
Model-χ^2 (df) P		466	(19)	.00000			
Likelihood Ratio Pseudo R^2 (McFadden)		.096					
Model-Fit: Hosmer-Lemeshow Test: [χ^2, (df), P]		10.3	(8)	.243			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). K: Kategoriale Variable: Die nicht explizit ausgewiesenen Kategorien (teils inkl. fehlende Werte) bilden die Referenzkategorie. 1) Zentriert und fehlende Werte (N=19) durch Mittelwert substituiert. 2) Ordinalvariable (Rating: st37q01), fehlende Werte (N=116) durch Mittelwert substituiert. 3) Ordinalvariable, fehlende Werte in Kategorie 1 (kein Mobiltelefon).

sierungsgrad sehr deutlich. Die Differenzen können prinzipiell auf einer unterschiedlichen Zusammensetzung der befragten kantonalen Teilsamples oder auf unbekannten raumbezogenen Kontextbedingungen beruhen. Da die Kantone vielerorts der intermediären Ebene der PISA-Erhebungsorganisation entsprechen, wäre es aber auch denkbar, dass sich hier ein verzögerter Einfluss der kantonal variablen Testadministration bemerkbar macht, der sich in einer mehr oder weniger positiven Einstellung der Jugendlichen gegenüber der TREE-Erhebung niederschlägt. Für diese Interpretation würde die diesmal recht deutliche Beteiligungsdiffe-

renz sprechen, die sich innerhalb des Wallis entlang der Sprachgrenze manifestiert.[1] Im Wallis gehören deutsch- und der französischsprachiger Kantonsteil nämlich zu zwei unterschiedlichen PISA-Erhebungsregionen (siehe auch Abschnitt 3.2).

Wendet man sich den *Individualmerkmalen* zu, so beteiligen sich SchülerInnen, die zur Zeit der PISA-Testsession noch *nicht im 9. Schuljahr* waren, seltener an der ersten Befragungswelle. Das Befragungsinstrument ist vermutlich weniger gut auf die oft atypischen Bildungsverläufe dieser eventuell auch schwerer erreichbaren Teilpopulation zugeschnitten, was die geringere Beteiligung erklären dürfte.[2] Die hohe Nonresponse-Quote wird dabei dadurch gemildert, dass sich genau diese Gruppe überdurchschnittlich an der Adressenerhebung beteiligt hatte (Tabelle 4). Einen starken Einfluss hat auch die *Lesekompetenz*, der sich nun erwartungsgemäss losgelöst von der Testadministration entfaltet. Jugendliche mit geringer Lesekompetenz nehmen dabei, wie zu befürchten war, erheblich weniger häufig an Welle 1 teil. Darüber hinaus hat auch die an der *Mathematiknote* gemessene Schulleistung einen Einfluss, wobei Jugendliche mit einer ungenügenden oder fehlenden Notenangabe die Teilnahme häufiger verweigern. Weiter hängt die Teilnahme auch von den *individuellen Plänen* fürs kommende Jahr ab. Weitaus am geringsten ist die Beteiligung dabei unter den SchülerInnen, die eine *Erwerbstätigkeit* anstreben oder die noch keinen bestimmten Plan haben. Leicht überdurchschnittlich oft nehmen dagegen Jugendliche teil, die ein *schulisches Zwischenjahr* einschalten wollen, was mit einem erhöhten Interesse am Befragungsthema gut erklärt werden kann.[3] Ein weiterer, ähnlich gelagerter Effekt geht von einer negativen *Einstellung zum Schulbesuch aus*. Hinzu kommt der Indikator für die *zeitgerechte Erledigung der Hausaufgaben*, welcher ebenfalls negativ mit der Beteiligung zusammenhängt und eine zusätzliche, mit der Einstellung zur Schule verwandte Einstellungsdimension abbildet. Neben den soweit genannten schulischen Grössen beinflusst in geringerem Masse auch der *kulturelle Hintergrund* der Befragten die Teilnahmebereitschaft. Jugendliche, die mindestens vier Mal jährlich eine *Rock-, Pop- oder Technoveranstaltung* besuchen, nehmen in geringerem Masse an der ersten Befragungswelle teil. Der Indikator verweist auf subkulturelle Zugehörigkeiten, die sich teilweise mit einer (verstärkten) Abgrenzung von der Erwachsenenwelt verbinden, was sich auch in der Beteiligung entsprechend niederschlägt. Die Zahl der *Mobiltelefone* und der Zahl der *Taschenrechner* im elterlichen Haushalt sind Indikatoren für je spezifische kulturelle Herkunftsmilieus im Bourdieuschen Sinne. Dabei scheint – ironischerweise – die Erreichbarkeit und damit auch die Teilnahme von Jugendlichen umgekehrt proportional zur Zahl der Mobiltelefone *ab*zunehmen. Umgekehrt liegt die Beteiligung von Jugendlichen aus Familien mit drei oder mehr Taschenrechnern deutlich höher, was sich wohl einem familiären Hintergrund zuschreiben lässt, in dem viel Wert auf Berechenbarkeit und Zuverlässigkeit gelegt wird. Vom familiären Hintergrund gehen davon losgelöst noch weitere Einflüsse aus. So haben Jugendliche, die nicht mit *Vater und/oder Mutter zusammenwohnen*, seltener an der Befragung teilgenommen. Weiter sind jene ausländischen Jugendlichen, die *ausserhalb Mitteleuropas geboren* sind, unter den Antwortenden ebenfalls unterproportional vertreten. Schliesslich haben sich junge *Frauen* wie schon bei der Adressenerhebung häufiger beteiligt.

---

[1] Fürs Unterwallis, das zur Romandie zählt, liegt das Verhältnis zwischen Teilnehmenden und Nicht-Teilnehmenden ceteris paribus nahe beim Durchschnitt (Exp[B] = Exp[.620-.707]=.92), fürs Oberwallis aber deutlich tiefer (Exp[-.456]=.63).

[2] Es handelt sich um Mitglieder des 15jährigen-Samples, die zur Zeit der PISA-Testsession noch eine siebte oder achte Klasse besucht haben.

[3] Die betreffende Kategorie umfasst zudem auch Jugendliche, die zwar bei der PISA-Befragung noch in die 8. oder 9. Klasse wechseln bzw. diese nachholen wollten, die aber de facto dennoch zur Population zählen, welche die Schule nach PISA verlassen hat.

3.4 Teilnahmewahrscheinlichkeit Welle 2

Das Vorgehen bei der Schätzung der individuellen Teilnahmewahrscheinlichkeiten für die zweite Befragungswelle entspricht im Wesentlichen demjenigen bei der ersten Welle. Zunächst wird geprüft, welche PISA-Merkmale für die Teilnahme an Welle 2 relevant sind. Anschliessend wird unter den Welle-1-Variablen systematisch nach weiteren Prädiktoren gesucht. Neben den diversen in Welle 1 erhobenen TREE-Skalen werden dabei, wo dies sinnvoll scheint, punktuell auch einzelne Items als Prädiktoren in Betracht gezogen. Die so ermittelten Effekte verbleiben im Modell, soweit sie theoretisch plausibel scheinen, statistisch ausreichend gesichert und robust sind.

Das zugrunde liegende Sample bilden all jene PISA-ProbandInnen, die das Adressblatt zurückgesandt haben und die nicht retrospektiv als nicht zur Grundgesamtheit zählend ausgeschlossen werden mussten (N=5943, siehe Tabelle 1). Als Teilnehmende gelten wie bei Welle 1 alle 5210 Samplemitglieder, die entweder den schriftlichen *Fragebogen* ausgefüllt haben oder mit denen ein telefonisches *Interview* realisiert worden ist.

Die *Kennwerte für das Gesamtmodell in Tabelle 5* weisen auf ein gut an die Daten angepasstes Regressionsmodell (Hosmer-Lemeshow-Test) sowie auf eine leicht stärkere systematische Variation der Nonresponse-Raten (McFaddenR^2) hin, als dies noch bei Welle 1 der Fall war. Gemessen an der Aufteilung der Likelihood-Beiträge leisten praktisch ausschliesslich das Antwortverhalten bei Welle 1 (≈ 81 χ^2-Punkte) und weitere Individualmerkmale (≈ 209 χ^2-Punkte) einen substanziellen Beitrag zur Modellanpassung. Der Beitrag des räumlichen Kontextes bleibt dagegen bescheiden (≈ 11 χ^2-Punkte).

Was den *räumlichen Kontext* betrifft, so fällt die Teilnahme im Kanton *Wallis* erneut geringer aus; Hinweise zur etwas spekulativen Interpretation dieses Effekts finden sich in Abschnitt 3.3. Davon einmal abgesehen, ist der räumliche Kontext nun aber nicht länger relevant.

Da TREE jenen Jugendlichen, die nicht bereit sind, den gesamte Fragenkatalog schriftlich zu beantworten, die Möglichkeit bietet, das Frageprogramm vollständig oder, falls auch dazu die Bereitschaft fehlt, in reduzierter Form telefonisch zu beantworten, gibt die Art der Beteiligung an der ersten Befragungswelle unmittelbar Aufschluss über deren Beteiligungs- bzw. Kooperationsbereitschaft. Sofern es sich dabei um ein über die Zeit relativ stabiles Individualmerkmal handelt, wird das beobachtete Teilnahmeverhalten somit einen guten Prädiktor für die spätere Beteiligung abgeben. Tatsächlich erweist sich das *Teilnahmeverhalten bei Welle 1* als der weitaus stärkste Prädiktor der Beteiligung an Welle 2. Etwas überraschend ist dabei, dass die Beteiligung unter den VerweigererInnen der Welle 1 nun leicht höher liegt als unter den Jugendlichen, die bei Welle 1 ein langes Telefoninterview absolviert haben.[1] Des weiteren erweist sich, dass ProbandInnen weniger oft teilnehmen, die bei Welle 1 den *Geburtsmonat* nicht angegeben haben – ein guter Indikator für ein rasches, unsorgfältiges Ausfüllen des Fragebogens und damit ebenfalls für eine reduzierte Antwortbereitschaft.

Seitens der direkt beobachteten Individualmerkmale erweist sich wiederum die *Lesekompetenz* als starker Prädiktor, wobei sich der stets gleichgerichtete Einfluss über die diversen Befragungsschritte hinweg natürlich zunehmend kumuliert. Häufiger nehmen zudem Lehrlinge mit einer hohen *Zufriedenheit mit den LehrmeisterInnen* teil. Ein sehr ausgeprägtes *Interesse der LehrmeisterInnen* an der Ausbildung der Befragten wirkt sich dagegen negativ

[1] Hingegen fallen Jugendliche, die bei T1 lediglich zu einem kurzen Telefoninterview bereit waren, nun nicht mehr gegenüber jenen Jugendlichen ab, die bei T1 das volle Frageprogramm schriftlich beantwortet haben.

Tabelle 5: *Schätzung Teilnahmewahrscheinlichkeit Befragungswelle 2*

Logistische Regression (N=5943)		B	S.E.	Wald	df	Sig.	Exp(B)
Räumlicher Kontext							
Kantone Wallis [D]	(canton = 23)	– .518	.152	11.6	1	.001	.596
Teilnahmeverhalten Welle 1							
Trackingstatus Welle 1 [K]				82.9	2	.000	
Telefoninterview lang	(track1 = 3)	– 1.180	.138	72.9	1	.000	.307
Nicht teilgenommen	(track1 = 1, 99)	– .833	.180	21.4	1	.000	.435
Geburtsmonat: Angabe fehlt [D]	(t1bim = 99)	– .529	.123	18.5	1	.000	.589
Individualmerkmale							
Lesekompetenz [1)]	(wleread)	.003	.001	40.8	1	.000	1.003
Zufriedenheit LehrmeisterIn [2)]	(t1quam1)	.285	.095	8.9	1	.003	1.330
LehrmeisterIn: Sehr interessiert [D]	(t1supi6 = 4)	– .417	.108	14.9	1	.000	.659
Ausbildungsstatus: Nicht in Ausb. [D]	(t1educ17 ≠ 5– 8, .)	– .301	.104	8.3	1	.004	.740
Hausaufgaben: immer fertig | Missing [D]	(st32q01= 4, 5)	.392	.105	14.0	1	.000	1.480
Skala: Copingverhalten (Avoiding) [3)]	(t1coa_3s)	– .215	.063	11.6	1	.001	.807
Skala: körperliche Beschwerden [3)]	(t1hea_8s)	– .213	.075	8.2	1	.004	.808
Skala: Suchtmittelkonsum [3)]	(t1dru_7s)	– .343	.114	9.0	1	.003	.709
Zuhause: Anz. Mobiltelefone (Rating) [4)]	(st22q01)	– .150	.042	12.9	1	.000	.861
Zuhause: Anzahl Taschenrechner (≥3) [D]	(st22q03 = 4)	.420	.098	18.3	1	.000	1.522
Haushaltstyp: WG, Anderes [D 5)]	(t1hous4 oder t1hous6 = 2)	– .593	.130	20.7	1	.000	.552
Lebt ohne Mutter und/oder Vater [D]	(st04q01 oder st04q03 = 2)	– .293	.103	8.1	1	.004	.746
Geburtsland: nicht Schweiz [D]	(st16q03 = 2)	– .300	.092	10.7	1	.001	.741
Geschlecht: Weiblich [D]	(st03q01 = 1)	.514	.093	30.6	1	.000	1.672
Regressionskonstante		1.473	.497	8.8	1	.003	4.363
Modellkennwerte							
– 2 Log-Likelihood		3859					
Model-χ^2 (df) P		580	(18)	.00000			
Likelihood-Ratio Pseudo R^2 (McFadden)		.130					
Model-Fit: Hosmer– Lemeshow Test: [χ^2, (df), P]		5.7	(8)	.676			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). K: Kategoriale Variable: Die nicht explizit ausgewiesenen Kategorien (teils inklusive fehlende Werte) bilden die Referenzkategorie. 1) Zentriert und fehlende Werte (N≤19) durch den Mittelwert substituiert. 2) Für die Fälle ohne Lehrmeister/in oder mit fehlender Angabe dazu durch den Mittelwert substitiert. 3) Fehlende Angaben durch Skalenmittelwert ersetzt. 4) Ordinalvariable, fehlende Werte in Kategorie 1 (kein Mobiltelefon). 5) Durchgängig fehlende Angaben zum Haushalt (t1hous1– t1hous6) sind ebenfalls mit 1 codiert.

aus; möglicherweise wird die Befragung durch TREE in dieser Situation als eine weitere unerwünschte Einmischung empfunden. Jugendliche, die zum Zeitpunkt der ersten Befragungswelle noch *keine anerkannte Ausbildung* – Berufslehre, Diplommittelschule, Berufsmatur oder Maturitätsschule – angefangen haben, nehmen ebenfalls seltener an Welle 2 teil. Der Effekt der *Einstellung zu Hausaufgaben* (PISA) pflanzt sich auch noch in der zweiten Welle fort, wobei nun aber − spiegelbildlich zum Effekt bei T1 − jene Jugendlichen öfter mitmachen, die ihre Aufgaben in der obligatorischen Schule *immer* rechtzeitig fertiggestellt hatten. Unter den Teilnehmenden unterproportional vertreten sind im Weiteren ProbandInnen, die, gemessen an den betreffenden Skalen, *Problemen ausweichen* (Coping-Skala), die oft *körperliche Beschwerden* haben und einen hohen *Suchtmittelkonsum* aufweisen. Des Weiteren pflanzen sich auch die *'kulturellen' Einflüsse* des familiären Hintergrundes weiterhin fort, die

sich an der Zahl der *Mobiltelefone* und der *Taschenrechner* im Haushalt festmachen. Wer in einer *Wohngemeinschaft*, einer anderen nicht-traditionellen Wohnform, oder in einer *'Broken Home'-Situation lebt*, nimmt ebenfalls unterdurchschnittlich oft an Welle 2 teil. Dies gilt auch für Jugendliche *ausländischer Herkunft*, und zwar – anders als bei Welle 1 – unabhängig davon, ob sie in einem mitteleuropäischen Land geboren worden sind. Und schliesslich bestätigt sich auch erneut, dass die jungen Frauen häufiger teilnehmen.

Bei der Schätzung des Modells stellt sich das Problem, dass die Informationen aus Welle 1 nur für jene Probandinnen verfügbar sind, die daran teilgenommen haben. Diese bilden zwar die grosse Mehrheit. In der Ausgangsstichprobe von Welle 2 finden sich aber auch gut 400 Jugendliche (≈7%), die sich nicht an Welle 1 beteiligt hatten. Von diesen haben sich rund zwei Drittel an Welle 2 beteiligt (!), was darauf hindeutet, dass der Welle-1-Nonresponse nur zum kleineren Teil auf 'harte', unwiderrufliche Verweigerungen zurückgeht. Die fehlenden Angaben zu den Welle-1-Prädiktoren im Modell werden bei den betroffenen ProbandInnen durch den Variablenmittelwert bzw. -modus substituiert.

Die Effektschätzungen der Welle-1-Variablen im Modell, die in der Tabelle am Präfix 't1…' der Variablennamen in den Klammern erkennbar sind, bleiben von der Substitution der fehlenden Werte praktisch unberührt, wie eine Vergleichsrechnung unter Ausschluss der betreffenden Jugendlichen zeigt. Die potentielle Verzerrung der resultierenden Gewichtungsvariable durch die Substitution ist somit ebenfalls als bescheiden einzuschätzen. Dafür sprechen einerseits die geringe Zahl und die vergleichsweise schwachen Effekte der involvierten Welle-1-Merkmale, andererseits die begrenzte Zahl an 'RückkehrerInnen' ins Welle-2-Sample (N = 288). In den Modellen für die weiteren Befragungswellen (Abschnitt 3.5-3.10) wird mit fehlenden Angaben aufgrund eines wellenspezifischen Unit-Nonresponse jeweils auf analoge Weise verfahren.

3.5 Teilnahmewahrscheinlichkeit Welle 3

Das Vorgehen bei der Schätzung der individuellen Teilnahmewahrscheinlichkeiten bei der dritten und allen weiteren Wellen entspricht demjenigen bei der zweiten Welle, soweit im jeweiligen Abschnitt nichts anderes erwähnt ist. Mit Blick auf Welle 3 wird zunächst ein Ausgangsmodell mit jenen PISA- und Welle-1-Variablen geschätzt, die sich bei der Erklärung der Beteiligung an Welle 1 und 2 als wichtig erwiesen haben.[1] Anschliessend werden all jene Prädiktoren, die sich gemessen an einem Likelihood-Ratio-Test ($\alpha \leq 1\%$) als unbedeutend erweisen, sukzessive wieder aus dem Modell entfernt. Da zwischen dem früheren Antwortverhalten und der Teilnahmebereitschaft bei Welle 3 starke Zusammenhänge zu vermuten sind, wird das Modell sodann um zwei Indikatoren erweitert, welche die Beteiligung an den beiden vorherigen Befragungswellen (Schriftlich, Telefon lang, Telefon kurz, Verweigert) erfassen, sowie um die Interaktion zwischen diesen. Schliesslich wird in den Daten der Welle 2 systematisch nach theoretisch plausiblen weiteren Prädiktoren gesucht. Das Resultat dieser Modellierungen ist in *Tabelle 6* dargestellt.

Wendet man sich zunächst den Kennwerten für das Gesamtmodell im unteren Teil der Tabelle zu, so zeigt sich, dass die Modellanpassung (Hosmer-Lemeshow-Test) und die Stärke des Zusammenhangs zwischen den verwendeten Prädiktoren und der Nonresponse-Rate (McFadden R^2) sich auf sehr ähnlichem Niveau bewegen wie schon beim Modell für die zweite Wel-

[1] Auch bei der Entwicklung der Modelle für alle weiteren Wellen werden jeweils die für die beiden vorherigen Wellen relevanten Prädiktoren berücksichtigt.

le. Dies spricht für alles in allem lediglich mässig starke Zusammenhänge zwischen Untersuchungsmerkmalen und der Teilnahme an Welle 3, zumal relativ umfassend nach Effekten von theoretisch potentiell rücklaufbestimmenden Untersuchungsmerkmalen gesucht worden ist. Anders als sonst üblich ist das eher bescheidene Pseudo-R^2 im Kontext der gegebenen Anwendung sehr erfreulich, da es auf einen relativ unsystematischen Ausfallprozess hinweist.

Tabelle 6: *Schätzung Teilnahmewahrscheinlichkeit Befragungswelle 3*

Logistische Regression (N=5599)		B	S.E.	Wald	df	Sig.	Exp(B)
Bisheriges Teilnahmeverhalten (Welle 1-2)							
T1: Telefoninterview | Verweigerung [D]	(t1track ≠ 2)	– .891	.115	60.2	1	.000	.410
T2: Telefoninterview lang [D]	(t2track = 3)	– 1.031	.177	34.1	1	.000	.357
Interaktionseffekt:							
T1: Tel.-Int. | Verweigerung * T2: Telefoninterview		.877	.218	16.1	1	.000	2.405
Geburtsmonat: Angabe fehlt [D]	(t1bim = 99)	– .536	.114	22.1	1	.000	.585
Individualmerkmale							
Lesekompetenz [1)]	(wleread)	.003	.001	31.2	1	.000	1.003
Schultyp: Sek., Real [D] (bereinigt)	(typ=2,3)	– .373	.099	14.3	1	.000	.689
Ausbildungsrealität: Nicht wie erwartet [D]	(t2expe1=2)	.385	.098	15.4	1	.000	1.470
Uninteressierte / keine Schulfreunde [D]	(t2supi4=1,5)	– .450	.138	10.7	1	.001	.638
Lehrbetriebswechsel [D]	(t2crit09=2,3)	– .598	.190	9.9	1	.002	.550
Jobwechsel [D]	(t2crit15=3,4)	– 1.181	.342	11.9	1	.001	.307
Auszug aus dem Elternhaus [D]	(t2clev02=2, 9, .)	– 1.038	.109	91.5	1	.000	.354
Skala: Suchtmittelkonsum [2)]	(t2drug_7s)	– .344	.102	11.4	1	.001	.709
Geschlecht: Weiblich (bereinigt) [D]	(st03q01 = 1)	.356	.086	16.9	1	.000	1.427
Regressionskonstante		1.397	.346	16.3	1	.000	4.042
Modellkennwerte							
– 2 Log-Likelihood		3730					
Model-χ^2 (df) P		562	(13)	.00000			
Likelihood-Ratio Pseudo R^2 (McFadden)		.131					
Model-Fit: Hosmer-Lemeshow Test: [χ^2, (df), P]		7.5	(8)	.481			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). 1) Zentriert und fehlende Werte durch den Mittelwert substituiert. 2) Fehlende Angaben sind durch den Skalenmittelwert ersetzt.

Betrachtet man die einzelnen Effektkoeffizienten des Modells, so ragen zunächst die sehr bedeutenden *Einflüsse des früheren Antwortverhaltens* hervor. Dies ist nicht allzu überraschend, da die Teilnahmebereitschaft intraindividuell eine gewisse Stabilität aufweisen dürfte. Die starken Effekte des Antwortverhaltens bei Welle 1 und 2 belegen unter Berücksichtigung der betreffenden Interaktion, dass insbesondere jene Jugendlichen weitaus häufiger an Welle 3 teilnehmen, die bei beiden früheren Wellen jeweils das volle Frageprogramm schriftlich beantwortet haben. Praktisch gleich hoch liegt die Teilnahmebereitschaft unter jenen Jugendlichen, die bei Welle 2 lediglich an einem *kurzen Telefoninterview* mit reduziertem Frageprogramm teilgenommen haben. Erheblich tiefer liegt die Teilnahme dagegen namentlich unter jenen Jugendlichen, mit denen bei Welle 1 *oder* 2 ein *langes Telefoninterview* geführt worden

ist.[1] Die langen Telefoninterviews haben sich demnach offenbar negativ auf die weitere Kooperationsbereitschaft ausgewirkt. Vermutlich werden die langen Telefoninterviews als ermüdend oder langweilig erlebt, was die weitere Kooperationsbereitschaft nachhaltig unterhöhlt. Diese Interpretation wird zum einen durch die Beobachtung gestützt, dass die *kurzen* Telefoninterviews offenbar keine negativen Spuren hinterlassen. Zum anderen ist auch bereits bei der Modellierung der Beteiligung an Welle 2 ein sehr ähnliches Effektmuster zutage getreten: Jugendliche, mit denen bei Welle 1 ein langes Telefoninterview geführt wurde, haben ceteris paribus massiv seltener an Welle 2 teilgenommen als die übrigen Teilnehmenden, und nur geringfügig häufiger als die Jugendlichen, die Welle 1 ganz verweigert hatten. (siehe oben, Tabelle 5).

Das Nennen des Geburtsmonats bei der T1-Befragung ist ein weiterer Indikator für die Antwortbereitschaft. Von diesem geht wie schon beim T2-Modell wiederum ein recht starker negativen Effekt aus, der sich dahingehend interpretieren lässt, dass die Teilnahmebereitschaft erneut bei jenen Jugendlichen geringer ist, die den T1-Fragebogen nur widerwillig und unsorgfältig ausgefüllt hatten.

Auch seitens der *Individualmerkmale* im Modell finden sich eine Reihe von 'alten Bekannten'. Dies trifft auf die Lesekompetenz, den Suchtmittelkonsum und auf das Geschlecht zu. Beim Indikator für den Suchtmittelkonsum handelt es sich nun aber neu um die T2-Messung. Was Richtung und Stärke der genannten Effekte angeht, so könnte man fast schon sagen, dass sie sich 'im gewohnten Rahmen' bewegen. Die über mehrere Wellen gleichgerichteten Einflüsse der 'alten Bekannten' weisen dabei unerfreulicherweise auf eine diesbezüglich zunehmend einseitige Zusammensetzung des Panel-Samples hin (siehe dazu Abschnitt 3.11).

Die Effekte der übrigen Prädiktoren lassen sich wie folgt interpretieren: Der Einfluss des *Schultyps* zeigt, dass Jugendliche, die zur Zeit der PISA-Erhebung eine Real- oder Sekundarschule besucht haben, sich weniger oft an Welle 3 beteiligt haben als Jugendliche, die damals in integrierten Schulen oder einem Gymnasium waren.

Die übrigen Prädiktoren entstammen der zweiten Befragungswelle. Jugendliche, deren Erwartungen nicht mit der angetroffenen *Ausbildungsrealität* übereinstimmen, haben häufiger an Welle 3 teilgenommen. Dies kann auf ein grösseres Interesse an den Befragungsthemen zurückgeführt werden. Wer keine oder an Ausbildungsfragen *nicht interessierte SchulkollegInnen* hat, nimmt dagegen weniger häufig teil, was auf eine kontext- bzw. klassenabhängige Sensibilisierung für die Wichtigkeit der Befragungsthematik zurückgeführt werden kann. Ein *Wechsel des Lehrbetriebs* und noch mehr ein *mindestens zweimaliges Wechseln des Jobs* noch vor der Zweitbefragung zieht eine deutlich tiefere Beteiligungsquote bei der Drittbefragung nach sich. Beides kann als Hinweis auf individuelle Schwierigkeiten beim Übergang an der ersten Schwelle gedeutet werden. Wer den *Auszug aus dem Elternhaus* frühzeitig vollzieht (oder wer keine Angabe darüber macht), nimmt ebenfalls wesentlich seltener an Welle 3 teil. Dafür können sowohl Schwierigkeiten, welche die neu gewonnene Eigenständigkeit mit sich bringen kann, als auch das Wegfallen der elterlichen Kontrolle verantwortlich sein.

[1] Die modellbasierte Teilnahmewahrscheinlichkeit für die ersten beiden erstgenannten Gruppen erreicht rund 92 Prozent, verglichen mit Werten um 80 Prozent für die letztgenannte (Schätzwerte für männliche, bezüglich Lesekompetenz, Suchtmittelkonsum usw. durchschnittliche Jugendliche). Eine ähnlich geringe erwartete Teilnahmewahrscheinlichkeit resultiert für Jugendliche, die sich nicht oder nur in reduzierter Form an Welle 1 beteiligt haben − es sei denn, sie sind bei Welle 2 für ein telefonisches Kurzinterview gewonnen worden.

3.6 Teilnahmewahrscheinlichkeit Welle 4

Das Modell für die Teilnahme an der vierten Welle in *Tabelle 7* zeigt wie gewohnt eine gute Anpassung an die Daten (Hosmer-Lemeshow-Test), wobei das höhere McFadden-R^2 allerdings auf einen insgesamt etwas stärkeren Zusammenhang zwischen Prädiktoren und Beteiligung hinweist, als dies bislang der Fall war.

Erstmals seit der ersten TREE-Welle finden sich wieder namhafte regionale Differenzen in den Teilnahmeraten, die diesmal für die Romandie und die italienischsprachige Schweiz deutlich höher liegen als für die deutschsprachige Referenzregion. Ab Welle 4 schwächt sich die aus der Adressenerhebung herrührende Unterrepräsentation der Romandie (exklusive Genf) wie bereits bei der ersten TREE-Befragungswelle nochmals deutlich ab. Hingegen akzentuiert sich die ebenfalls auf die Adressenerhebung zurückgehende Übervertretung von Jugendlichen aus dem italienischsprachigen Landesteil mit der vierten Welle noch merklich.

Tabelle 7: *Schätzung Teilnahmewahrscheinlichkeit Befragungswelle 4*

Logistische Regression (N=5333)		B	S.E.	Wald	df	Sig.	Exp(B)
Räumlicher Kontext							
Romandie D	(reg_5 = 3,4)	.779	.161	23.5	1	.000	2.179
Italienischsprachige Schweiz D	(reg_5 = 5)	.698	.103	45.6	1	.000	2.010
Bisheriges Teilnahmeverhalten (Welle 1-3)							
T1: Telefoninterview, Verweigerung D	(t1track ≠ 2)	– .405	.119	11.6	1	.001	.667
T2: Telefoninterview, Verweigerung D	(t2track ≠ 2)	– .514	.120	18.2	1	.000	.598
T3: Trackingstatus K				151.1	2	.000	
Telefoninterview	(t3track = 3, 4)	– 1.816	.148	150.9	1	.000	.163
Nicht teilgenommen	(t3track = 1)	– .907	.134	45.9	1	.000	.404
Interaktionseffekt:							
T2: Telefoninterview * T3: Nicht teilgenommen		.788	.259	9.3	1	.002	2.199
Individualmerkmale							
Lesekompetenz 1)	(wleread)	.003	.001	27.2	1	.000	1.003
Zuhause: Anzahl Mobiltelefone (Rating) 2)	(st22q01)	– .158	.045	12.5	1	.000	.853
Zukunftspläne: Mit Ausb. weiterfahren D	(t3plan1=1)	.630	.137	21.2	1	.000	1.878
Zukunftspläne: Andere / neue Ausb. D	(t3plan3=1)	.548	.187	8.6	1	.003	1.729
Skala: Arbeitsorganisatorische Probleme 3)	(t2jorg_2s)	.738	.221	11.1	1	.001	2.093
Wohnform: Lebt mit (Ehe-)PartnerIn D	(t3hous7 = 1)	– .742	.257	8.3	1	.004	.476
Regressionskonstante		– .383	.534	0.5	1	.473	.681
Modellkennwerte							
– 2 Log-Likelihood		3233					
Model-χ^2 (df) P		732	(13)	.00000			
Likelihood-Ratio Pseudo R^2 (McFadden)		.185					
Model-Fit: Hosmer-Lemeshow Test: [χ^2, (df), P]		8.2	(8)	.412			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). K: Kategoriale Variable: Die nicht explizit ausgewiesenen Kategorien (teils inklusive fehlende Werte) bilden die Referenzkategorie. 1) Zentriert und fehlende Werte durch den Mittelwert substituiert. 2) Ordinalvariable, fehlende Werte in Kategorie 1 (kein Handy). 3) Fehlende Angaben sind durch den Skalenmittelwert ersetzt.

Die bedeutendsten Einflüsse gehen wiederum von den Indikatoren für das frühere *Teilnahmeverhalten* aus. Die Befunde zeigen, dass Jugendliche, die bei einer der bisherigen drei Wellen entweder gar nicht oder dann bloss in Form eines Telefoninterviews teilgenommen haben, sich erheblich seltener an Welle 4 beteiligen. Die betreffenden negativen Effekte schwächen sich dabei um so mehr ab, je weiter die fragliche Welle zurück liegt. Mit Blick auf die Beteiligung an Welle 3, von der die weitaus stärksten Effekte ausgehen, wirkt sich dabei ein Telefoninterview noch weit negativer aus als eine Nicht-Teilnahme. Dies kann wohl auf die im letzten Abschnitt skizzierten 'Treatment'-Effekt von langen Telefoninterviews zurückgeführt werden. Der positive Interaktionseffekt zwischen der Teilnahme T2 und T3 gibt zugleich einen Hinweis, dass sich dieser Treatment-Effekt nach einer einmaligen Nicht-Teilnahme wieder verflüchtigt.

Mit Blick auf die *Individualmerkmale* erweist sich einmal mehr die von PISA erfasste *Lesekompetenz* als bedeutend. Da sich deren Einflüsse über die Wellen kumulieren, setzt sich das Panel unerfreulicherweise zunehmend einseitiger aus Jugendlichen einer anfänglich guten Lesefähigkeit zusammen. Zum wiederholten Male zeigt sich zudem ein negativer Zusammenhang zwischen der Beteiligung und der *Zahl der Mobiltelefone* im elterlichen Haushalt (PISA-Messung*)*, die als Indikator für überdurchschnittlich mobile, schwer erreichbare Jugendliche bzw. Familien gelesen werden kann. Jugendliche mit klaren *Zukunftsplänen* zum weiteren Ausbildungsverlauf haben sich häufiger beteiligt als andere Jugendliche. Dasselbe gilt für Jugendliche innerhalb von *Betrieben* mit subjektiv als ungünstig erlebten *Arbeitsbedingungen* (arbeitsorganisatorische Probleme). Beides dürfte mit einer erhöhten Sensibilisierung für die Wichtigkeit der TREE-Befragungsthemen einhergehen, die sich positiv auf die Beteiligung auswirkt. Schliesslich nehmen jene Jugendlichen merklich seltener teil, die frühzeitig aus dem Elternhaus ausgezogen sind und die bereits *mit Partner oder Partnerin zusammenwohnen*.

3.7 Teilnahmewahrscheinlichkeit Welle 5

Das in *Tabelle 8* dargestellte Modell für Welle 5 ist gemäss Hosmer-Lemeshow-Test recht gut an die Daten angepasst, während der Zusammenhang zwischen Prädiktoren und Beteiligung gemäss McFadden-R^2 etwas schwächer ausfällt als noch bei Welle 4.

Verglichen mit der letzten Welle spielt der *räumliche Kontext* für die Teilnahme an Welle 5 wieder eine relativ bescheidene Rolle. Jugendliche aus der Romandie haben nun wieder *unter*durchschnittlich partizipiert, wie der Effekt der betreffenden dichotomen Variablen belegt, womit sich die noch aus der Adressenerhebung herrührende Unterrepräsentation in der Stichprobe wieder etwas akzentuiert.

Erneut kommt dem früheren *Teilnahmeverhalten* eine Schlüsselrolle zu. Jugendliche, die die Teilnahme an mindestens einer der beiden letzten Wellen verweigert haben, nehmen nun auch weitaus seltener an Welle 5 teil. Eine Nicht-Teilnahme an der unmittelbar vorangehenden Welle erhöht das Risiko einer erneuten Verweigerung dabei deutlich stärker. Was die vorletzte Welle angeht, so beteiligen sich die damals telefonisch Befragten ebenso selten wie die damaligen Verweigerer. Dieses Ergebnis liegt auf der Linie der nun mehrfach replizierten Befunde zu einem ungünstigen 'Treatment'-Effekt der Telefoninterviews. Mit Blick auf die *Individualmerkmale* erweist sich die PISA-*Lesekompetenz* unverändert als bedeutend. Die diesbezüglich zunehmend einseitige Zusammensetzung der Panelstichprobe verstärkt sich somit unerfreulicherweise erneut. Jugendliche, deren ausbildungsbezogene Erwartungen bei der Zweitbefragung nicht mit der *Ausbildungsrealität* übereinstimmen, nehmen − wie schon

Tabelle 8: *Schätzung Teilnahmewahrscheinlichkeit Befragungswelle 5*

Logistische Regression (N=5049)		B	S.E.	Wald	df	Sig.	Exp(B)
Räumlicher Kontext							
Romandie [D]	(reg_5 = 3,4)	– .315	.101	9.7	1	.002	.730
Bisheriges Teilnahmeverhalten (Welle 1-4)							
T3: Telefoninterview | Verweigerung [D]	(t3track ≠ 2)	– .576	.111	26.9	1	.000	.562
T4: Verweigerung [D]	(t4track = 1, 15)	– 1.077	.145	55.0	1	.000	.340
Individualmerkmale							
Lesekompetenz [1)]	(wleread)	.002	.001	10.5	1	.001	1.002
Ausbildungsrealität: Nicht wie erwartet [D]	(t2expe1=2)	.382	.102	13.9	1	.000	1.466
Zukunftspläne: Mit Ausb. weiterfahren [D]	(t4plan1 = 1)	.673	.144	21.9	1	.000	1.959
Zukunftspläne: Andere / neue Ausb. [D]	(t4plan3 = 1)	.460	.135	11.7	1	.001	1.585
Selbstbeteiligung Ausbildungskosten [K]				59.1	2	.000	
gar keine Selbstbeteiligung	(t4moex4 = 1)	– .533	.150	12.6	1	.000	.587
Angabe fehlt	(t4moex4 = -1, 9)	– 1.128	.147	58.5	1	.000	.324
Wohnortswechsel Familie [D]	(t4clev1 = 2,8)	– .493	.176	7.9	1	.005	.611
Vater- / Mutterschaft [D]	(t4clev11 = 2,8)	– 1.377	.418	10.8	1	.001	.252
Regressionskonstante		1.879	.323	33.8	1	.000	6.545
Modellkennwerte							
– 2 Log-Likelihood		2868					
Model-χ^2 (df) P		588	(11)	.00000			
Likelihood-Ratio Pseudo R^2 (McFadden)		.154					
Model-Fit: Hosmer-Lemeshow Test: [χ^2, (df), P]		12.3	(8)	.140			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). K: Kategoriale Variable: Die nicht explizit ausgewiesenen Kategorien (teils inklusive fehlende Werte) bilden die Referenzkategorie. 1) Zentriert und fehlende Werte durch den Mittelwert substituiert.

bei Welle 3 − wiederum besonders häufig teil. Dasselbe gilt zudem für Jugendliche, die bei der letzten Befragungswelle klare *Zukunftspläne* zum weiteren Ausbildungsverlauf geäussert haben. Beides dürfte mit einer erhöhten Sensibilisierung für die Bedeutung der TREE-Befragungsthemen einhergehen. Als bedeutender Prädiktor der Teilnahme erweist sich auch die *Selbstbeteiligung der Jugendlichen an den Ausbildungskosten*: Jugendliche, die finanziell nichts beisteuern und noch mehr solche, die die betreffende Angabe verweigern, nehmen seltener an Welle 5 teil. Auch dieser Effekt kann als Ausdruck einer mit der Selbstbeteiligung zunehmenden Sensibilisierung für die Bedeutung von Ausbildungsfragen und -entscheidungen interpretiert werden. Schliesslich wirken sich zwei der erfassten 'Critical Life Events' negativ auf die Beteiligung im Folgejahr aus, nämlich ein *Wohnortswechsel* der Familie sowie eine frühe *Vater- oder Mutterschaft*.

3.8 Teilnahmewahrscheinlichkeit Welle 6

Das Modell zur Schätzung der Teilnahmewahrscheinlichkeiten für Welle 6 ist in *Tabelle 9* dargestellt. Wie die bisherigen Modelle ist es gut an die Daten angepasst (Hosmer-Lemeshow-Test). Das McFadden-R^2 weist allerdings auf einen insgesamt relativ starken Zusammenhang zwischen Prädiktoren und Beteiligung hin, wie er bislang einzig für das Modell zur Adressenerhebung zu verzeichnen war.

Tabelle 9: *Schätzung Teilnahmewahrscheinlichkeit Befragungswelle 6*

Logistische Regression (N=4844)		B	S.E.	Wald	df	Sig.	Exp(B)
Bisheriges Teilnahmeverhalten (Welle 1-5)							
T3: Verweigerung [D]	(t4track = 1)	– .536	.136	15.5	1	.000	.585
T4: Verweigerung [D]	(t4track = 1, 15)	– .994	.141	49.5	1	.000	.370
T5: Trackingstatus [K]				189.3	2	.000	
Nur CATI | Basis- & Erg.-Fragebogen	(t5track = 2, 3, 5)	– 1.059	.115	84.8	1	.000	.347
Verweigerung | Nur Basisfragebogen	(t5track = 4, 9)	– 1.939	.160	146.5	1	.000	.144
Individualmerkmale							
Berufs(Matur) | Mittelschulabschluss [D]	(t5crtyp.. = 3-6, 8)	.651	.111	34.5	1	.000	1.917
Zukunftspläne: Mit Ausb. weiterfahren [D]	(t3plan1=1)	– .324	.102	10.0	1	.002	.724
Selbstwertgefühl: Sehr gering, Missing [D]	(t5sele4s < 3.5 | .)	– .579	.148	15.3	1	.000	.560
Auszug aus Elternhaus [D]	(t5clev02 = 2,8)	– .435	.151	8.3	1	.004	.647
Regressionskonstante		2.764	.103	720.6	1	.000	15.862
Modellkennwerte							
– 2 Log-Likelihood		3199					
Model-χ^2 (df) P		835	(8)	.00000			
Likelihood-Ratio Pseudo R^2 (McFadden)		.207					
Model-Fit: Hosmer-Lemeshow Test: [χ^2, (df), P]		8.3	(7)	.309			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). K: Kategoriale Variable: Die nicht explizit ausgewiesenen Kategorien (teils inklusive fehlende Werte) bilden die Referenzkategorie.

Womöglich noch mehr als bisher erweist sich das *frühere Teilnahmeverhalten* als Schlüsselgrösse zur Erklärung der aktuellen Teilnahme. Dabei nehmen namentlich Jugendliche, die bei mindestens einer der drei letzten Wellen die Teilnahme verweigert haben, nun auch seltener an Welle 6 teil. Wie nicht anders zu erwarten, schwächen sich die betreffenden Effekte dabei mit wachsender zeitlicher Distanz zusehends ab − am stärksten ist der Effekt einer Nicht-Teilnahme an der unmittelbar vorangehenden Welle. Eine deutlich reduzierte Beteiligung findet sich ferner auch bei allen Befragten der fünften Welle, die damals auf einen der beiden alternativen Befragungsmodi (siehe Tabelle 2) ausgewichen sind.[1]

Mit Blick auf die *Individualmerkmale* hat die PISA-*Lesekompetenz* erfreulicherweise erstmals keinen Effekt mehr. Auch sonst erweisen sich nur wenige Individualmerkmale als bedeutsam. Deutlich öfter haben sich Jugendliche beteiligt, die bis Welle 5 einen *Maturitäts- oder Mittelschulabschluss* erlangt hatten. Dies ist insofern einleuchtend, als dieser relativ sprach- und

[1] Die Beteiligung der kleinen Gruppe, die bei Welle 5 bloss den Basisfragebogen ausgefüllt hatte, liegt nun ceteris paribus sogar ebenso tief wie unter den damaligen Verweigerern.

schriftgewandten Gruppe eine Teilnahme leichter fällt; auch eine höhere Motivation respektive ein grösseres Interesse an der Befragungsthematik mag eine Rolle spielen. Die Überrepräsentation von Jugendlichen mit klaren, auf die Beendigung der Ausbildung zielenden *Zukunftsplänen* im Panel, die sich über die beiden letzten Wellen aufgebaut hat, schwächt sich aufgrund des nunmehr negativen Effektes tendenziell etwas ab. Deutlich tiefer liegt ferner auch die Beteiligung von Jugendlichen, die entweder ein geringes *Selbstwertgefühl* mitbringen oder die betreffenden Fragen nicht beantwortet haben. Und schliesslich beeinflusst erneut auch einer der erfassten 'Critical Life Events' die Beteiligung im Folgejahr, nämlich der *Auszug aus dem Elternhaus*.

3.9 Teilnahmewahrscheinlichkeit Welle 7

Das Schätzmodell für Welle 7 ist in *Tabelle 10* dargestellt. Wie gewohnt ist es gut an die Daten angepasst (Hosmer-Lemeshow-Test), wobei das McFadden-R^2 nun allerdings sogar noch etwas höher liegt als bei Welle 6 und beim Adressblattmodell. Dies weist auf einen besonders engen Zusammenhang zwischen Prädiktoren und Beteiligung hin.

Tabelle 10: *Schätzung Teilnahmewahrscheinlichkeit Befragungswelle 7*

Logistische Regression (N=4640)		B	S.E.	Wald	df	Sig.	Exp(B)
Räumlicher Kontext							
Genf (Erhebungsregion) [D]	(reg_5 = 4)	– .423	.132	10.2	1	.001	.655
Bisheriges Teilnahmeverhalten (Welle 1-6)							
T4: Verweigerung [D]	(t4track = 1)	– .526	.173	9.3	1	.002	.591
T5: Trackingstatus [K]				91.7	2	.000	
Nur CATI | Basis- & Erg.-Fragebogen	(t5track = 2, 3, 5)	– .722	.123	34.2	1	.000	.486
Verweigerung | Nur Basisfragebogen	(t5track = 4, 7)	– 1.321	.142	86.8	1	.000	.267
T6: Trackingstatus [K]				174.6	2	.000	
Nur CATI	(t6track = 3)	– 1.220	.153	63.7	1	.000	.295
Verweigerung | Nur Basisfragebogen	(t6track = 4, 7)	– 2.059	.156	174.3	1	.000	.128
T4: Kostenbet. Ausbildung: Angabe fehlt	(t4moex4 = -1, 9)	– .399	.120	11.0	1	.001	.671
Individualmerkmale							
Berufsausbildung abgeschlossen [D]	(t6credu1. ≠ 96002, .)	.289	.109	7.1	1	.008	1.335
Zukunftspläne: Mit Ausb. weiterfahren [D]	(t6plan1=1)	.867	.145	35.9	1	.000	2.379
Zukunftspläne: Andere / neue Ausb. [D]	(t6plan3=1)	.539	.174	9.6	1	.002	1.715
Lebt nicht mit Vater [D]	(t6hous2 = 1)	– .593	.159	14.0	1	.000	.552
Geschlecht: Weiblich (bereinigt) [D]	(st03q01 = 1)	– .395	.105	14.2	1	.000	.674
Regressionskonstante		3.110	.170	336.1	1	.000	22.420
Modellkennwerte							
– 2 Log-Likelihood		2753					
Model-χ^2 (df) P		1036	(12)	.00000			
Likelihood-Ratio Pseudo R^2 (McFadden)		.273					
Model-Fit: Hosmer-Lemeshow Test: [χ^2, (df), P]		8.4	(8)	.391			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). K: Kategoriale Variable: Die nicht explizit ausgewiesenen Kategorien (teils inklusive fehlende Werte) bilden die Referenzkategorie.

Nachdem der *räumliche Kontext* bei der sechsten Welle keine Rolle gespielt hat, fällt die Beteiligung der Jugendlichen aus der Erhebungsregion Genf bei Welle 7 geringer aus. Der Effekt ist eher schwierig zu interpretieren, aber robust und gut gesichert. Bei Genf handelt es sich um die einzige Erhebungsregion, für die der TREE-Längsschnitt auf einem annähernd kompletten Adressensatz aufbaut (Abschnitt 3.2), was dann bei der ersten TREE-Welle zunächst zu einer entsprechend deutlich verminderten Teilnahme geführt hat (Tabelle 4). Bei Welle 4 und 5 ist in der gesamten Romandie inklusive Genf zuerst eine ceteris paribus deutlich höhere und unmittelbar darauf eine etwas geringere Teilnahme zu registrieren (Tabelle 7 und 8). Aufkumuliert über die Adressenerhebung und die ersten sechs Wellen ist für Genf indessen immer noch von einer leicht überdurchschnittlichen Ausschöpfung der Population auszugehen, womit sich diese aufgrund des nunmehr dem negativen Effekts dem schweizerischen Durchschnitt angleicht.

Erneut gehen die weitaus bedeutendsten Effekte vom *früheren Teilnahmeverhalten* aus, das über die damalige Antwortbereitschaft Aufschluss gibt. Wie schon bei Welle 6 nehmen namentlich Jugendliche seltener teil, die bei mindestens einer der drei letzten Wellen die Teilnahme verweigert haben. Und wiederum schwächen sich die betreffenden Effekte mit wachsender zeitlicher Distanz zusehends ab − am stärksten ist der Effekt einer Nicht-Teilnahme an der unmittelbar vorangehenden sechsten Welle.[1] Deutlich seltener nehmen ferner Jugendliche teil, die sich bei Welle 5 oder 6 lediglich zu einem CATI-Interview bereit erklärt haben.[2] Die Befunde zur Teilnahmebereitschaft liegen auf der Linie der bisherigen Befunde. Wie schon bei Welle 5 erweist sich zudem die *Auskunftsbereitschaft* der Befragten hinsichtlich ihrer *Selbstbeteiligung an den Ausbildungskosten* als wichtig; allerdings beschränkt sich der Effekt nunmehr auf den Unterschied zwischen antwortbereiten und nicht auskunftswilligen Befragten. Anders als noch im T5-Modell kann er daher vollumfänglich auf eine entsprechend geringere Teilnahmebereitschaft der nicht Auskunftswilligen zurückgeführt werden.

Die Befunde zu den *Individualmerkmalen* bestätigen zunächst die Beobachtung von Welle 6, wonach die PISA-*Lesekompetenz* ihre Bedeutung eingebüsst hat. Dieser positive Befund dürfte sich nicht zuletzt der partiellen Umstellung der Erhebung auf CATI verdanken. Auch sonst erweisen sich wiederum nur relativ wenige Individualmerkmale als bedeutsam. Jugendliche, die bis Welle 6 eine *Berufsausbildung abgeschlossen* haben, nehmen öfter teil als andere. Damit zeigt sich nun für die Berufsbildungen (inkl. schulische) um ein Jahr verzögert ein sehr ähnlicher Effekt wie er bei Welle 6 für die Mittelschulabschlüsse zu beobachten war. Dies leuchtet insofern ein, als sich nach jedem Ausbildungsabschluss die Frage nach dem 'Wie weiter?' stellt, was das Interesse der Jugendlichen an den TREE-Kernthemen steigert. Deutlich höher liegt die Beteiligung zudem einmal mehr unter Jugendlichen mit klaren, auf die Fortsetzung oder aber einen Wechsel der Ausbildung zielenden *Zukunftsplänen*. Wie schon bei Welle 1 und 2 beteiligen sich zudem Jugendliche erheblich seltener, die *nicht mit dem Vater zusammenleben*. Eine Überraschung birgt der Befund zum *Geschlecht*: Junge Frauen beteiligen sich nämlich − ein absolutes Novum − erstmals weniger oft als junge Männer. Die sich über die diversen Wellen nach und nach akzentuierende Überrepräsentation der Frauen im Sample schwächt sich damit erstmals wieder etwas ab.

---

[1] Wie schon bei Welle 6 liegt die Beteiligung der kleinen Gruppe, die zuvor bloss den Basisfragebogen ausgefüllt hat, ceteris paribus ebenso tief wie unter den damaligen Verweigerern.

[2] Diese Kategorie schliesst mit Blick auf die Teilnahme an Welle 5 zudem auch eine kleinere Anzahl an Jugendlichen ein, die das CATI abgebrochen oder Basis- und Ergänzungsbogen ausgefüllt haben.

3.10 Teilnahmewahrscheinlichkeit Welle 8

Die achte Befragungswelle ist – nach zwei Jahren ohne weitere Befragung – im Jahr 2010 realisiert worden, als die Kohortenmitglieder zirka 26 Jahre alt waren. Das Schätzmodell für die Teilnahmewahrscheinlichkeit an der achten Welle ist in *Tabelle 11* dargestellt. Das Modell zeigt eine akzeptable Anpassung an die Daten (Hosmer-Lemeshow-Test) und es weist auf einen gegenüber Welle 7 nun wieder merklich schwächeren Zusammenhang zwischen Prädiktoren und Teilnahme hin (McFadden-$R^{2)}$.

Tabelle 11: *Schätzung Teilnahmewahrscheinlichkeit Befragungswelle 8*

Logistische Regression (N=4505) [1)]		B	S.E.	Wald	df	Sig.	Exp(B)
Bisheriges Teilnahmeverhalten (Welle 1-7)							
T4: Verweigerung [D]	(t4track = 1)	-.735	.158	21.6	1	.000	.479
T5: Reduzierte Beteiligung [D]	(t5track ≠ 1)	-.489	.093	27.6	1	.000	.613
T6: Reduzierte Beteiligung [D]	(t6track = 3,4,7)	-.593	.097	37.2	1	.000	.553
T7: Trackingstatus [K]				227.6	3	.000	
Basis--&Ergänzungs-Fragebogen	(t7track = 2)	-.475	.179	7.0	1	.008	.622
nur CATI	(t7track = 3,5)	-.828	.109	58.2	1	.000	.437
Verweigerung	(t7track = 7)	-1.833	.122	226.2	1	.000	.160
Individualmerkmale							
Lesekompetenz [2)]	(wleread)	.003	.000	43.4	1	.000	1.003
Selbstwertgefühl: Sehr gering, Missing [D]	(t5sele4s < 3.5 | .)	-.392	.110	12.6	1	.000	.676
Skala: Bewältigung Emotionszentriert [2)]	(t7cope4s)	.191	.072	7.1	1	.008	1.211
Skala: Bewältigung Vermeidungszentriert [3)]	(t7cope4s)	-.221	.074	8.8	1	.003	.802
Regressionskonstante		.649	.362	3.2	1	.073	1.913
Modellkennwerte							
– 2 Log-Likelihood		3927					
Model-χ^2 (df) P		1040	(10)	.00000			
Likelihood-Ratio Pseudo R^2 (McFadden)		.210					
Model-Fit: Hosmer-Lemeshow Test: [χ^2, (df), P]		12.3	(8)	.139			

D: Dichotome Variable (Eins, wenn der Klammerausdruck wahr ist, sonst Null). K: Kategoriale Variable: Die nicht explizit ausgewiesenen Kategorien (teils inklusive fehlende Werte) bilden die Referenzkategorie. 1) Ein Fall mit nachträglich eingetroffenem Basis-Fragebogen bei der Modellierung irrtümlich als Nicht-Antwortend codiert (ID 69005009). 2) Zentriert und fehlende Werte durch den Mittelwert substituiert. 2) Fehlende Angaben sind durch den Skalenmittelwert ersetzt.

Das bereits bestens bekannte Ergebnis, wonach das *frühere Teilnahmeverhalten* den besten Prädikor für die Teilnahmebereitschaft und damit auch für das aktuelle Teilnahmeverhalten abgibt, zeigt sich ungeachtet des grösseren zeitlichen Abstands wiederum auch bei der achten Befragungswelle: Wer in der vierten, fünften, sechsten oder siebten Befragungswelle entweder gar nicht oder dann nur in einer aufwandreduzierten Form teilgenommen hat, beteiligt sich mit stark erhöhter Wahrscheinlichkeit auch nicht an Welle 8.

Für den räumlich-bildungsinstitutionellen Kontext sowie für *individuelle Merkmale der Kohortenmitglieder* – die primären Untersuchungsmerkmale – lassen sich dagegen gar keine respektive nur vergleichsweise schwache Effekte nachweisen. Im einzelnen zeigt sich nun erstmals seit Welle 5 wieder ein Einfluss der *PISA-Lesekompetenz* am Ende der neunten Klasse, die die Beteiligung wiederum positiv beeinflusst. Die bereits erhebliche diesbezügliche Selek-

tivität der Panelstichprobe verstärkt sich damit leider erneut. Bereits von Welle 6 bekannt auch die ungünstige Wirkung eines sehr *geringem Selbstwertgefühl*s auf die Teilnahmewahrscheinlichkeit. Weitgehend neu (siehe aber Welle 2, Tabelle 5) sind hingegen die beiden Einflüsse des individuellen Bewältigungsverhaltens: Hier zeigt sich, dass ein vermeidungszentriertes Verhalten die Teilnahmewahrscheinlichkeit senkt, während ein emotionszentriertes 'Coping' sie erhöht.

3.11 Kumulative Wirkungen des Nonresponse

Mit den im letzten Abschnitt geschätzten Teilnahmewahrscheinlichkeiten an der achten TREE-Folgebefragung $A_{W8.i}$ liegen nun für sämtliche Wellen die benötigten Bausteine zur Berechnung von Individualgewichten gemäss Beziehung (1) vor. Bevor in den folgenden Abschnitten auf die Stutzung, nachträgliche Schichtung und Kalibrierung der so gewonnenen Rohgewichte eingegangen wird, soll noch kurz die Frage nach den kumulativen, sich über die verstärkenden oder kompensierenden Wirkungen des Nonresponse auf die Zusammensetzung der jeweils noch verbleibenden Panelstichprobe aufgegriffen werden.

Grundsätzlich ist davon auszugehen, dass die individuellen Unterschiede in den Einschlusswahrscheinlichkeiten ins Panelsample einer gegebenen Welle teils auf das PISA-Stichprobendesign und teils auf das systematisch variierende individuelle Teilnahmeverhalten zurück gehen. Das relative Gewicht des Teilnahmeverhaltens gegenüber den designbedingten Unterschieden erhöht sich dabei von Welle zu Welle. Dementsprechend sinkt der Anteil der rein designbedingten Unterschiede in den individuellen Einschlusswahrscheinlichkeiten ins Panel (Kehrwert der Gewichte gemäss Beziehung 1) bis zur achten Welle von 61 auf 37 Prozent ab. Die ausschliesslich durch das Stichprobendesign bedingten Unterschiede sind dabei nicht weiter problematisch, da sie von Spezifikations- und Schätzfehlern frei sind. Die Modelle zur Korrektur der systematischen Unterschiede im Teilnahmeverhalten bilden die zu Grunde liegenden Selektionsprozesse hingegen bestenfalls in guter Näherung ab. So gesehen ist es sehr zu begrüssen, dass die Unterschiede in den individuellen Einschlusswahrscheinlichkeiten zu einem sehr grossen, wenn auch erwartungsgemäss von Welle zu Welle schrumpfenden Teil auf das stark disproportionale PISA-Erhebungsdesign zurückgeführt werden können. Deren Variabilität fusst mit anderen Worten wesentlich auf den je nach Teilstichprobe stark *unterschiedlichen Auswahlsätzen*. So gesehen korrigieren die konstruierten Gewichtungen zu erheblichen Teilen einfach die weit überproportionale Vertretung der Romandie innerhalb der Ausgangsstichprobe.

Mit Blick auf die Zusammensetzung des *Nonresponse* ist dabei zunächst von Interesse, wie sich der Zusammenhang zwischen den thematisch im Zentrum von TREE stehenden Untersuchungsmerkmalen und der Teilnahmebereitschaft über die Wellen hinweg verändert hat. Zu diesen Untersuchungsmerkmalen im engeren Sinne zählen dabei alle Individualmerkmale, nicht aber die Indikatoren für die Antwortbereitschaft (bisheriges Teilnahmeverhalten), die Testadministration und die Kontextvariablen. Die folgende Tabelle enthält dazu zwei Kennzahlen für jede Welle, nämlich je ein McFadden-R^2 für das vollständige Modell jeder Welle sowie für ein entsprechendes Modell *ohne* Individualmerkmale. Der Vergleich der beiden Kennwerte gibt Aufschluss darüber, welchen Beitrag die Untersuchungsmerkmale im engeren Sinne zur Modellanpassung beisteuern.

Vor allem bei den ersten drei Wellen und bei der letzten beruht die Modellanpassung massgeblich auf den einbezogenen Individualmerkmalen. Insbesondere bei diesen Wellen ist der Nonresponse also mit Blick auf relevante Untersuchungsmerkmale nicht neutral zusammen-

gesetzt. Was die ersten drei Wellen angeht, wird der bedeutende *relative* Beitrag der Individualmerkmale allerdings durch den − wie das McFadden-R^2 für das vollständige Modell anzeigt − *insgesamt* relativ bescheidenen Zusammenhang zwischen Prädiktoren und Teilnahme relativiert. Offensichtlich bloss eine marginale Rolle spielen die Individualmerkmale in den Modellen zu den Wellen 4 bis 6. Da unter den TREE-Untersuchungsmerkmalen jeweils relativ umfassend nach potenziell relevanten Prädiktoren der Teilnahme gesucht worden ist, legen diese Befunde insgesamt den Schluss nahe, dass der Nonresponse bezogen auf die eigentlichen Untersuchungsmerkmale bis Welle 6 über weite Strecken unsystematisch zusammengesetzt war, was mit Blick auf dessen potenziell verzerrende Wirkung sehr erfreulich ist. Bei der siebten Befragungswelle hängt die Teilnahme dann allerdings stärker von diversen relevanten Individualmerkmalen ab (Tabelle 10), so dass der Bias hier potenziell einiges grösser ausfällt. Bei der achten TREE-Welle hat sich die Zusammensetzung der Panelstichprobe dann glücklicherweise wieder nur noch wenig verändert.

Tabelle 12: *Beitrag der Individualmerkmale zu den wellenspezifischen Modellen*

	Likelihood-Ratio Pseudo R^2 (McFadden)	
	vollständiges Modell	reduziertes Modell [1)]
Adressblatt-Rücklauf	.202	.140
Welle 1	.096	.032
Welle 2	.130	.001
Welle 3	.131	.048
Welle 4	.185	.161
Welle 5	.154	.122
Welle 6	.207	.190
Welle 7	.391	.254
Welle 8	.210	.204

1) Reduzierte Modelle gemäss Tabellen 3 bis 11 ohne Individualmerkmale (und zugeh. Interaktionen).

Allerdings findet sich bei einer detaillierten Betrachtung der einzelnen Teilnahmemodelle (Tabelle 3 bis 11) eine Reihe von Individualmerkmalen, die die Beteiligung über mehrere Wellen hinweg beeinflussen, so dass sich deren Wirkungen auf die Samplezusammensetzung im Zeitverlauf entsprechend kumulieren. Um solche *kumulativen Wirkungen* auf die Zusammensetzung der Panelstichprobe einzuschätzen, wird zunächst auf Basis der Modelle in Tabelle 3 bis 11 die Wahrscheinlichkeit errechnet, dass Jugendlichen mit einem bestimmten teilnahmerelevanten Merkmal bis zu einer gegebenen Befragungswelle in der Panelstichprobe verbleiben.[1] In *Tabelle 13* sind diese Wahrscheinlichkeiten in Relation gesetzt zur entsprechenden Wahrscheinlichkeit für 'durchschnittliche' Jugendliche ohne das betreffende Merkmal.[2] Die dargestellte *'relative Verbleibswahrscheinlichkeit'* gibt somit auf anschauliche Wei-

[1] Dafür werden zunächst anhand von Beziehung (4) die wellenspezifischen Teilnahmewahrscheinlichkeiten für − ansonsten 'durchschnittliche' − Jugendliche mit dem jeweiligen Merkmal berechnet und dann über die Wellen ausmultipliziert, um die bedingte Wahrscheinlichkeit ihres Verbleibs im Panel bis Welle x zu erhalten.

[2] Für *'durchschnittliche' Jugendliche* wird die Wahrscheinlichkeit eines Verbleibs im Panel bis zur Welle x ebenfalls durch Ausmultiplizieren der Teilnahmewahrscheinlichkeiten der ersten x Wellen ermittelt. Bei Skalen und graduell abgestuften Prädiktoren wird den 'Durchschnittsjugendlichen' dabei jeweils der Variablenmittelwert zugeordnet, bei kategorialen und dichotomen Prädiktoren zählen sie zur Referenzkategorie.

se darüber Aufschluss, um wieviel sich die individuelle Wahrscheinlichkeit des Verbleibs in der Panelstichprobe aufgrund des jeweiligen Merkmals erhöht oder vermindert.

Beispielsweise geht aus den ersten beiden Zeilen der Tabelle hervor, dass Jugendliche mit einer sehr hohen *Lesekompetenz* sehr viel häufiger im Panel verbleiben, solche mit einer geringen hingegen weitaus seltener als 'durchschnittliche Jugendliche'. Bereits bei der Adresserhebung liegt die Verbleibswahrscheinlichkeit der ersteren um 24 Prozent über dem Durchschnitt und diejenige der zweiten um fast ebensoviel darunter. Bis zur fünften Welle wird dann die Zusammensetzung des Panels bezüglich Lesekompetenz sukzessive noch einseitiger, so dass die relative Verbleibsquote der beiden betrachteten Gruppen bei den letzten drei

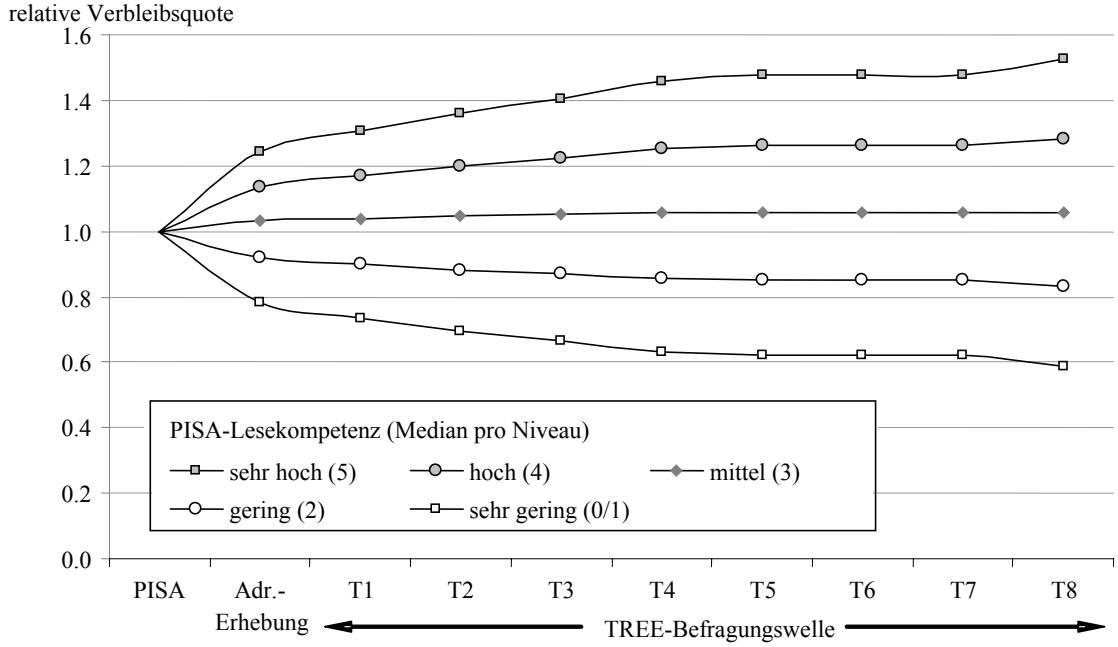
Tabelle 13: *Kumulative Wirkung ausgewählter Prädiktoren auf die 'Verbleibsquote' im Panel* *

Relative Verbleibswahrscheinlichkeit [1]		TREE-Befragungswelle								
Teilnahmerelevante Merkmale [2]		Adr. [4]	T1	T2 [5]	T3	T4	T5	T6	T7	T8
Lesekompetenz PISA [3]	sehr hoch	1.24	1.31	1.36	1.40	1.46	1.48	1.48	1.48	1.53
	sehr gering	0.78	0.73	0.69	0.67	0.63	0.62	0.62	0.62	0.58
Nicht im 9. Schuljahr		1.20	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
Zukunftspläne: Mit Ausb. weiterfahren		1.00	1.00	1.00	1.00	1.05	1.08	1.08	1.11	1.11
Zukunftspläne: Andere / neue Ausb.		1.00	1.00	1.00	1.00	1.05	1.07	1.07	1.09	1.09
Ausbildungsrealität: Nicht wie erwartet		1.00	1.00	1.00	1.03	1.03	1.05	1.05	1.05	1.05
Hausaufgaben rechtzeitig fertig: Nie		1.00	0.92	0.88	0.88	0.88	0.88	0.88	0.88	0.88
Skala Suchtmittelkonsum	sehr hoch	1.00	1.00	0.98	0.96	0.96	0.96	0.96	0.96	0.96
Skala Selbstwertgefühl	sehr gering	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.96	0.92
Zuhause: Anzahl Mobiltelefone (≥3)		1.00	0.97	0.95	0.95	0.92	0.92	0.92	0.92	0.92
Zuhause: Anzahl Taschenrechner (≥3)		1.00	1.04	1.07	1.07	1.07	1.07	1.07	1.07	1.07
Lebt nicht mit Mutter		1.00	0.92	0.89	0.89	0.89	0.89	0.89	0.89	0.89
Lebt nicht mit Vater		1.00	0.96	0.93	0.93	0.93	0.93	0.93	0.90	0.90
Geburtsland ausserhalb Mitteleuropa		0.92	0.86	0.83	0.83	0.83	0.83	0.83	0.83	0.83
Geschlecht: Weiblich		1.16	1.21	1.25	1.28	1.28	1.28	1.28	1.25	1.25

* Kleine Fehler gegenüber der Version von 2008 sind korrigiert. 1) Verhältnis zwischen der Wahrscheinlichkeit des Verbleibens eines Jugendlichen mit dem jeweiligen Merkmal in der Panelstichprobe und der entsprechenden Wahrscheinlichkeit eines 'durchschnittlichen' Jugendlichen (siehe Text). 2) Merkmale, die sich in den Modellen zu mindestens *zwei* Befragungswellen als relevant erweisen. 3) 'Sehr hoch' respektive 'sehr gering' entspricht dabei dem Medianwert der Jugendlichen auf PISA-Lesekompetenzniveau fünf respektive Eins oder Null in der TREE-Ausgangsstichprobe (wleread = 650 bzw. 359). 4) Der Berechnung der Teilnahmewahrscheinlichkeit an der Adresserhebung wird ein mittlerer Interaktionseffekt zwischen Lesekompetenz und Testadministration zugrunde gelegt (Cluster 9 gemäss Tabelle 3). 5) Bei der Berechnung der Teilnahmewahrscheinlichkeit zu Welle 2 dient − anders als in Tabelle 5 − 'Hausaufgaben: immer fertig | missing' als Referenzkategorie der betreffenden Variable.

Wellen schliesslich 53 Prozent über respektive um 42 Prozent unter dem Durchschnitt liegen. Die bei den ersten Wellen stets ähnlich starken, gleichgerichteten Effekte der Lesekompetenz kumulieren sich mithin in erheblichem Masse über die Wellen. Die folgende Grafik veranschaulicht das Ergebnis anhand der fünf PISA-Kompetenzniveaus (PISA 2002: 24f.).

Grafik 3: *Verbleib im Panel nach PISA-Lesekompetenzniveau*



Die Grafik verdeutlicht zum einen, dass die Verbleibsquote von Jugendlichen mit einer sehr hohen Lesekompetenz (PISA-Niveau 5) ab der fünften Befragungswelle fast zweieinhalb mal höher liegt als bei Jugendlichen mit einer sehr geringen Lesekompetenz (Niveau ≤ 1).[1] Kumuliert über die Befragungswellen geht von der Lesekompetenz also ein überaus starker Effekt auf den Verbleib in der Panelstichprobe aus. Zum anderen wird gleichzeitig auch deutlich, dass rund die Hälfte der Unterschiede in den Verbleibsquoten bereits auf die Adressblatt-Erhebung gleich zu Anfang zurückgeht. Die anfänglichen Unterschiede haben sich dann bis zur fünften Welle sukzessive weiter verstärkt, um sich dann auf hohem Niveau zu stabilisieren. Möglicherweise hat die bei der fünften Welle vorgenommene Umstellung des TREE-Basismoduls auf CATI (vgl. Tabelle 2) zu dieser erfreulichen Stabilisierung beigetragen.

Wendet man sich den verbleibenden Befunden in Tabelle 13 zu, so umfasst diese neben der Lesekompetenz entsprechende Auswertungen für sämtliche Merkmale, die in mindestens zwei der involvierten Modelle zur Erklärung der weiteren Teilnahme eine Rolle spielen und bei denen demzufolge *potenziell* mit einer Kumulation der Wirkungen zu rechnen ist. Im Unterschied zur Lesekompetenz bleiben die kumulativen Wirkungen allerdings mit wenigen Ausnahmen relativ bescheiden. Für die Mehrzahl der Merkmale bewegen sich die relativen Verbleibsraten in Tabelle 13 über sämtliche Wellen hinweg in einem Bereich zwischen rund 0.9 und 1.1, womit die kumulierte Abweichung von der durchschnittlichen Verbleibswahrscheinlichkeit maximal etwa 10 Prozent erreicht.

Stärkere Abweichungen sind allerdings *zum einen* für das Herkunftsland und das Geschlecht (ganz unten in der Tabelle) zu verzeichnen. Für die nicht aus einem mitteleuropäischen Land stammende grosse Mehrheit der *Jugendlichen ausländischer Herkunft* sinkt die Verbleibswahrscheinlichkeit bis Welle 2 auf 83 Prozent des Vergleichswerts. Anschliessend bleibt sie ceteris paribus konstant. *Junge Frauen* nehmen dagegen bereits weit häufiger an der Adress-

[1] Ab Welle 5 liegt das Verhältnis der jeweiligen Verbleibsquoten konstant bei etwa 2.4 (1.48 / 0.62).

erhebung teil (+16%), um anschliessend ihre Überrepräsentation im Panel bis zur dritten Welle auf 28 Prozent auszubauen. In den letzten beiden Wellen hat sich diese wieder leicht abgebaut (auf noch + 25%). *Zum anderen* kommen zu den negativen Wirkungen eines frühzeitiges Nicht-Zusammenlebens mit Mutter und/oder Vater (*je* etwa minus zehn Prozent, vgl. Tabelle 13) noch etliche punktuelle, wellenspezifische Effekte von Familienkonstellation, Wohnumfeld und 'Critical Life Events' hinzu, die in eine ähnliche Richtung weisen. So zeitigen das Wohnen in einer Wohngemeinschaft (T2) oder im Konkubinat (T4), ein früher Auszug aus dem Elternhaus (T3, T6) sowie eine frühe Elternschaft (T5) allesamt negative Effekte auf den Verbleib im Panel. Insgesamt sprechen diese Befunde dafür, dass Jugendliche aus unvollständigen Familien sowie solche, die frühzeitig 'flügge' werden, im Panel stärker unterrepräsentiert sind, als allein aufgrund von Tabelle 13 zu vermuten wäre.

Zusammenfassend lässt sich feststellen, dass sich die Zusammensetzung der Panelstichprobe im Befragungsverlauf primär bezüglich vier Individualmerkmalen verändert hat. Auf der einen Seite sind Jugendlichen mit geringer Lesekompetenz, junge Männer sowie ausländische Jugendliche bereits bei der Adressen-Erhebung überdurchschnittlich häufig aus der Stichprobe ausgeschieden. Deren erhebliche Unterrepräsentation bereits in der bereinigten TREE-Ausgangsstichprobe hat sich dann über die ersten Befragungswellen hinweg noch zusehends akzentuiert. Zum anderen sind Jugendliche aus vollständigen Familien, die im gesamten erfassten Lebensabschnitt im elterlichen Haushalt verbleiben, in der Panelstichprobe merklich überrepräsentiert.

4 Stutzung der Rohgewichte

Mit der Schätzung der Teilnahmewahrscheinlichkeiten für sämtliche Wellen liegen nun alle Elemente für die Berechnung von entsprechenden Gewichtungsvariablen gemäss Beziehung (1) vor. Unter der Annahme, dass die dem Nonresponse zu Grunde liegenden systematischen Ausfallprozesse durch die Modelle in guter Näherung abgebildet werden, lassen sich mit einer so konstruierten Gewichtung erwartungstreue Stichprobenschätzungen gewinnen.

Häufig besteht bei der Anwendung solcher Stichprobengewichtungen indes ein *Zielkonflikt* zwischen der angestrebten *Korrektur von Nonresponse-Verzerrungen* und der Minimierung der negativen Auswirkungen von Gewichtungen auf die *Präzision der samplebasierten Schätzungen* und Hochrechnungen. Dabei nimmt die zu erwartende Präzisionseinbusse grundsätzlich mit der Varianz der Gewichtungsvariablen zu. Gerade bei Panelgewichtungen schwillt die Streuung der Gewichte von Welle zu Welle unweigerlich an (siehe Beziehung 1), was die Schätzpräzision je länger je mehr beeinträchtigt. Dies gilt ganz speziell, wenn die Zahl der Wiederholungsbefragungen wie im Falle von TREE gross ist. Vielfach verhält es sich dabei so, dass nur einige wenige Fälle mit Extremwerten die Streuung der Gewichtungsvariablen sehr stark in die Höhe treiben, wodurch sich die Schätzgenauigkeit entsprechend deutlich vermindert. Nach Kish (1992) vergrössert sich etwa die Varianz einer gewichteten Mittelwertschätzung (μ_w) im Vergleich zu einer ungewichteten (μ) gemäss folgendem Ausdruck, wobei *cv* dem Variationskoeffizienten der verwendeten Gewichtungsvariablen entspricht.

$$\mathrm{var}(\mu_w) = \mathrm{var}(\mu) \cdot \left(1 + cv^2\right) \tag{5}$$

Neben einer der mehr oder weniger gravierenden Verminderung der Schätzpräzision haben extrem grosse Individualgewichte ausserdem auch den bedeutenden Nachteil, dass sie Auswertungen für kleinere Teilstichproben weit über Gebühr beeinflussen respektive unstabil machen können.

Die genannten Nachteile einer Gewichtung können mit einer *Stutzung der Gewichte* vermieden oder zumindest gemildert werden. Dabei werden alle Individualgewichte, die einen definierten oberen Eckwert übersteigen, auf diesen zurückgesetzt. Der *optimale Eckwert* für die Stutzung wird anhand einer Auswertung auf der Basis von Beziehung (5) bestimmt, wie sie in *Tabelle 14* beispielhaft für die T4-Gewichte dargestellt ist. Die im Beispiel verwendete Gewichtungsvariable gemäss Beziehung (1) ist dabei auf einen Mittelwert von 1 rekalibriert worden.[1] In der ersten Spalte ist der zu Probezwecken systematisch variierte obere Eckwert ausgewiesen, auf den die Gewichte gestutzt werden. In der folgenden Spalte sind die resultierenden Variationskoeffizienten der unterschiedlich gestutzten T4-Individualgewichte sowie deren Quadrat ausgewiesen. Aus der vierten Spalte geht hervor, wie sich die Varianz des gewichteten Stichprobenschätzers gemäss Beziehung (5) in Abhängigkeit vom gewählten Stutzungs-Eckwert vergrössert. Ohne jede Stutzung wäre somit mit einer rein gewichtungsbedingten Verminderung der Schätzgenauigkeit um etwa das Sechsfache zu rechnen. Je stärker die Variabilität der Gewichtungsvariablen mittels Stutzung vermindert wird, umso mehr sinkt dieses ungünstige Verhältnis dann nach und nach ab. Bezieht man zusätzlich auch die Zahl der von der Stutzung betroffenen Individualgewichte in der Spalte ganz rechts ein, so erweist sich im gewählten Beispiel eine Stutzung bei einem Eckwert von 8 als insgesamt optimal: Bei einer radikaleren Stutzung nimmt die Präzision der Sampleschätzung nur noch langsam zu, während zugleich die Zahl der betroffenen Gewichte bzw. ProbandInnen relativ rasch ansteigt, womit sich die Wirksamkeit der Nonresponse-Korrekturen und die Erwartungstreue der Schätzer zusehends vermindert. Es bestehen Hinweise, wonach der Trade-Off zwischen Präzisionsgewinn und Vergrösserung des Bias nur bei einer vorsichtigen, sich auf einen kleinen Anteil von Fällen beschränkenden Stutzung positiv ausfällt. Die im Beispiel betrachteten rekalibrierten T4-Rohgewichte werden aufgrund dieser Überlegungen bei einem oberen Eckwert von 8 gestutzt.[2] Wie aus der Tabelle hervorgeht, kann die gewichtungs

Tabelle 14: *Stutzen der T4-Rohgewichte*

	cv	var(μ_w)/var(μ)	Anzahl
ohne Stutzung	2.35	6.52	0
Stutzung der rekalibrierten G_i ab ...			
> 50	2.12	5.49	2
> 20	1.67	3.79	10
> 10	1.37	2.88	33
> 9	1.32	2.76	36
> 8	1.28	2.64	39
> 7	1.23	2.52	51
> 6	1.18	2.39	67
> 5	1.11	2.24	93

[1] Die rohe T1-Gewichtung wird dafür durch ihren Mittelwert dividiert. Die Rekalibrierung hat keinerlei Auswirkung auf die Optimierung der Stutzung.

[2] Umgerechnet auf den rohen Hochrechnungsfaktor gemäss Beziehung (1) entspricht dies einem oberen Eckwert von ungefähr 110.

bedingte Präzisionseinbusse dank der Stutzung von lediglich 39 Extremgewichten (0,8 % des T4-Samples) substanziell verringert werden: Anstatt sechseinhalbmal grösser sind die gemäss Beziehung (5) zu erwartenden Samplefehler dank der Stutzung nur noch gut zweieinhalb mal so gross wie bei einem *ungewichteten* Sample gleicher Grösse.

Auch bezogen auf die rohen Stichprobengewichtungen für die weiteren TREE-Wellen kann die Schätzgenauigkeit dank einer Stutzung massgeblich verbessert werden. Aufgrund der oben beispielhaft dargestellten Kriterien erweist sich für die Gewichtungen zu den ersten fünf Wellen jeweils eine Stutzung auf einen oberen Eckwert von acht als ideal, bei der sechsten und siebten TREE-Welle einer von sieben und bei der achten Welle einer von vier.[1] Aus der folgenden Tabelle geht hervor, wie sich die Schätzpräzision dank der Stutzung verbessert und wie viele Extremgewichte davon jeweils betroffen sind.

Tabelle 15: *Stutzung,Schätzpräzision und betroffene Fälle*

	ohne Stutzung	mit Stutzung	betroffene Gewichte	
	$var(\mu_w)/var(\mu)$	$var(\mu_w)/var(\mu)$	Anzahl	(%)
Welle 1-Sample	2.6	2.1	18	(0,3)
Welle 2-Sample	2.8	2.2	23	(0,4)
Welle 3-Sample	8.4	2.4	26	(0,5)
Welle 4-Sample	6.5	2.6	39	(0,8)
Welle 5-Sample	7.2	3.2	52	(1,2)
Welle 6-Sample	55.2	3.5	52	(1,3)
Welle 7-Sample	118.6	4.9	54	(1,4)
Welle 8-Sample	199.9	5.2	51	(1,5)

Alle ausgewiesenen Kennwerte auf Basis der pro Welle realisierten Samples.

Die Resultate in *Tabelle 15* machen zum einen deutlich, dass sich die Schätzpräzision bei einem Verzicht auf eine Stutzung ab Welle 3 und dann vor allem ab Welle 6 in einem nicht akzeptablen Mass verschlechtert. Sieht man von Analysen einmal ab, die ausschliesslich Daten der ersten beiden Wellen heranziehen, so ist aufgrund dieser Ergebnisse von Auswertungen mit ungestutzten Gewichten unbedingt abzuraten. Selbst mit der vorgenommenen Stutzung vermindert sich die Schätzpräzision gewichtungsbedingt noch erheblich und, wie aufgrund der von Welle zu Welle anwachsenden Streuung der Gewichte nicht anders zu erwarten ist, nimmt sie über den Längsschnitt hinweg nach und nach deutlich ab. Beachtlich ist vor allem die zusätzliche Genauigkeitseinbusse ab der siebten Welle. Bei der Interpretation der obigen Kennzahlen zur Schätzpräzision gilt es allerdings auch zu bedenken, dass es sich jeweils um *relative* Einbussen gegenüber einem ungewichteten Sample *gleichen Umfangs* handelt. Als Vergleichsmassstab dient somit stets eine äusserst präzise Schätzung, ist doch die TREE-Stichprobe aussergewöhnlich umfangreich. Selbst das Sample der achten Welle umfasst immer noch gut 3'400 Befragte, ein Stichprobenumfang, der weit genauere Schätzungen ermöglicht, als sonst vielfach üblich.

[1] Die genannten Stutzungspunkte beziehen sich weiterhin auf die Rohgewichte G_i gemäss Beziehung (1), die auf einen Mittelwert von 1 rekalibriert worden sind. Bei der Festsetzung des idealen Stutzungspunkts besteht stets ein erheblicher Ermessensspielraum; der Gewichtungsdatensatz umfasst aus diesem Grund auch ungestutzte Rohgewichte, was es erlaubt, den Trade-Off zwischen Präzision und Erwartungstreue je nach Analyse individuell festzusetzen.

5 Nachträgliche Schichtung

Die entwickelte Gewichtung soll unter anderem für die Abschätzung des absoluten und relativen Umfangs ausgewählter Teilpopulationen im Längsschnitt herangezogen werden. Mit Blick auf solche Populationshochrechnungen wird der Umfang einer Reihe besonders relevanter Teilpopulationen mittels einer nachträglichen Schichtung ('Poststratification', siehe Elliot 1992; Kish 1995) der konstruierten Gewichtungen über alle Befragungswellen hinweg konstant gehalten.[1] Dies kann zur Stabilisierung entsprechender Hochrechnungen beitragen.

Da es für die interessierende Schulabgänger-Population des Jahres 2000 an geeigneten Referenzverteilungen z. B. aus offiziellen Statistiken mangelt, stützt sich die nachträgliche Schichtung dabei auf die Welle-1-Daten. Diese sind nicht von der Panelmortalität betroffen und daher am besten geeignet, um die unbekannt bleibende Verteilung in der zu Grunde liegenden Population bestmöglich anzunähern. Die nachträgliche Schichtung berücksichtigt den auf Sekundarstufe I besuchten Schultyp, das Geschlecht und die Sprachregion, um gemäss der folgenden Tabelle zwölf Strata abzugrenzen, deren Umfang im Längsschnitt konstant gehalten wird:

Tabelle 16: *Referenzverteilung für die nachträgliche Stichprobenschichtung*

Schultyp Sek. I	Geschlecht	Sprachgebiet	Anteil (%)
erweiterte Ansprüche [1)]	weiblich	deutsch	24.9
erweiterte Ansprüche [1)]	weiblich	französisch	9.3
erweiterte Ansprüche [1)]	männlich	deutsch	22.1
erweiterte Ansprüche [1)]	männlich	französisch	8.8
Grundansprüche [2)]	weiblich	deutsch	10.1
Grundansprüche [2)]	weiblich	französisch	2.3
Grundansprüche [2)]	männlich	deutsch	13.9
Grundansprüche [2)]	männlich	französisch	2.3
integrierter Typ	weiblich	italienisch [3)]	1.1
integrierter Typ	weiblich	deutsch / französisch	1.8
integrierter Typ	männlich	italienisch [3)]	1.4
integrierter Typ	männlich	deutsch / französisch	2.2
Total			100.0

1) Sekundarschule oder Gymnasium. 2) Realschule. 3) Nur integrierter Typ vorhanden.

Die nachträgliche Schichtung sorgt dafür, dass die entsprechend rekalibrierten Stichprobengewichte aller späteren Wellen stets die in der Tabelle ausgewiesene Referenzverteilung auf Basis des Welle-1-Samples einhalten.[2]

[1] Sofern das Gewichtungsmodell nicht *sämtliche* Quellen systematischen Nonresponses in vollständig korrekter Spezifikation einschliesst, kann die gewichtete Sampleverteilung von der Ausgangsverteilung abweichen.

[2] Dafür werden die wellenspezifischen Gewichte jeweils mit einem stratumsspezifischen Kalibrierungsfaktor multipliziert. Zwei Fälle mit fehlendem Schultyp bleiben von der nachträglichen Schichtung ausgespart.

6 Hochrechnungsfaktoren und inferenzstatistische Gewichte

Anhand von Beziehung (1) berechnete Individualgewichte eignen sich nicht für inferenzstatistische Zwecke, sondern lediglich für *Hochrechnungen* auf die zu Grunde liegende Population (siehe Abschnitt 1). Eine mit G_i gewichtete Auszählung liefert einen Schätzwert für die Gesamtzahl der Populationsmitglieder mit den ausgezählten Merkmalen. Sobald *inferenzstatistische Auswertungen* ins Auge gefasst werden, bei denen Signifikanztests, Standardfehler und Vertrauensintervalle ins Spiel kommen, muss G_i aber jeweils so rekalibriert werden, dass die Summe der Gewichte dem Umfang des jeweils analysierten Samples entspricht (Moser & Kalton, 1972). Manche Statistikprogramme nehmen eine entsprechende Rekalibrierung automatisch vor, bei anderen, muss diese manuell durchgeführt werden.[1] Für jede TREE-Welle enthält der Datensatz mit den Gewichtungsvariablen neben Hochrechnungsgewichten deshalb auch entsprechend rekalibrierte inferenzstatistische Gewichte. Dabei gilt es allerdings zu beachten, dass für inferenzstatistische Auswertungen, die auf *Teilsamples* beruhen oder bei denen das Sample aufgrund von *fehlenden Werten* zusammenschrumpft, die vorgegebene Rekalibrierung jeweils neu anzupassen ist.[2] Der *Datensatz mit den TREE-Gewichten* umfasst − neben der PISA-Basisgewichtung gemäss Abschnitt 3.1 − je vier Varianten von Gewichtungsvariablen für jede TREE-Welle. Neben einem rohen Hochrechnungsgewicht gemäss Beziehung (1) ist auch ein gestutztes Hochrechnungsgewicht mit einer nachträglichen Schichtung gemäss Abschnitt 5 verfügbar. Die gestutzten und geschichteten Hochrechnungsgewichte sind dabei jeweils so rekalibriert, dass der *hochgerechnete Umfang der Gesamtpopulation* stets 80'000 beträgt. Dies entspricht näherungsweise dem nicht genau bekannten Umfang der zu Grunde liegenden Population.[3] Zu jedem der beiden Hochrechnungsgewichte umfasst der Datensatz zudem auch ein entsprechend aufgebautes inferenzstatistisches Gewicht, das sich einzig durch die Rekalibrierung auf einen Samplemittelwert von 1 davon unterscheidet. In der Regel werden für Auswertungszwecke am besten die gestutzten und kalibrierten inferenzstatistischen Gewichte herangezogen (siehe aber auch Abschnitt 7); die entsprechenden Hochrechungsgewichte kommen einzig zur Anwendung, wenn der *absolute* Umfang von interessierenden Teilpopulationen geschätzt werden soll.

In *Tabelle 17* sind die wichtigsten *Verteilungskennwerte der soweit gewonnenen Individualgewichte* für Hochrechnungen und inferenzstatistische Zwecke zusammengestellt. Die zu Grunde liegenden Samples bilden die jeweils an einer Welle teilnehmenden Jugendlichen (siehe auch Tabelle 1).[4]

---

1 Dafür wird das Hochrechnungsgewicht durch dessen Mittelwert im jeweils analysierten Sample dividiert.

2 Der *Mittelwert* der verwendeten Gewichtungsvariablen *im Analysesample* sollte stets Eins betragen.

3 Da nachträgliche Schichtung und Rekalibrierung *nach* der Stutzung der Gewichtungsvariablen erfolgen, können die Maximalwerte der kalibrierten Hochrechnungs- und Stichprobengewichte höher liegen, als die im letzten Abschnitt ausgewiesenen Stutzungskriterien.

4 Die aufgeführten Variablen sind allesamt im Datenfile mit den Gewichtungsvariablen (TREE-Weights_T1-T8.sav) enthalten. Dieses enthält zudem sämtliche 'Predicted Probabilities' zu den Modellen in Tabelle 3 bis 11 sowie die für Berechnung und nachträgliche Schichtung der Gewichte verwendeten Hilfsvariablen.

Tabelle 17: *Deskriptive Statistiken der Gewichtungsvariablen* [1)]

	Mittelwert	Summe	Standardab-weichung	Minimum	Maximum	N
Gewichtungen zu Welle 1						
roher Hochrechnungsfaktor	13.9	76620	17.8	1.36	691	5532
rohe inferenzstatistische Gewichtung	1.0	5532	1.3	0.10	50	5532
gestutzter & kalibrierter Hochrechnungsfaktor	14.5	80000	15.3	1.44	117	5532
gestutzte & kalibrierte Gewichtung	1.0	5532	1.1	0.10	8	5532
Gewichtungen zu Welle 2						
roher Hochrechnungsfaktor	16.0	83262	21.4	1.39	463	5210
rohe inferenzstatistische Gewichtung	1.0	5210	1.3	0.09	29	5210
gestutzter & kalibrierter Hochrechnungsfaktor	15.4	80000	16.9	1.30	145	5210
gestutzte & kalibrierte Gewichtung	1.0	5210	1.1	0.08	9	5210
Gewichtungen zu Welle 3						
roher Hochrechnungsfaktor	19.3	94166	52.4	1.42	2769	4882
rohe inferenzstatistische Gewichtung	1.0	4884	2.7	0.07	144	4882
gestutzter & kalibrierter Hochrechnungsfaktor	16.4	80069	19.4	1.16	147	4882
gestutzte & kalibrierte Gewichtung	1.0	4884	1.2	0.07	9	4882
Gewichtungen zu Welle 4						
roher Hochrechnungsfaktor	22.5	105290	52.9	1.45	1954	4680
rohe inferenzstatistische Gewichtung	1.0	4680	2.4	0.06	87	4680
gestutzter & kalibrierter Hochrechnungsfaktor	17.1	80000	22.3	1.14	177	4680
gestutzte & kalibrierte Gewichtung	1.0	4680	1.3	0.07	10	4680
Gewichtungen zu Welle 5						
roher Hochrechnungsfaktor	27.9	125779	69.3	1.48	1527	4504
rohe inferenzstatistische Gewichtung	1.0	4504	2.5	0.05	55	4504
gestutzter & kalibrierter Hochrechnungsfaktor	17.8	80000	23.7	0.74	201	4504
gestutzte & kalibrierte Gewichtung	1.0	4504	1.3	0.04	11	4504
Gewichtungen zu Welle 6						
roher Hochrechnungsfaktor	43.5	180026	320.4	1.55	15564	4135
rohe inferenzstatistische Gewichtung	1.0	4135	7.4	0.04	357	4135
gestutzter & kalibrierter Hochrechnungsfaktor	19.3	80000	28.5	0.86	262	4135
gestutzte & kalibrierte Gewichtung	1.0	4135	1.5	0.04	14	4135
Gewichtungen zu Welle 7						
roher Hochrechnungsfaktor	72.7	289341	788.0	1.62	35713	3982
rohe inferenzstatistische Gewichtung	1.0	3982	10.8	0.02	491	3982
gestutzter & kalibrierter Hochrechnungsfaktor	20.1	80000	35.9	0.47	404	3982
gestutzte & kalibrierte Gewichtung	1.0	3982	1.8	0.02	20	3982
Gewichtungen zu Welle 8						
roher Hochrechnungsfaktor	163.9	561171	2312.1	1.69	97089	3424
rohe inferenzstatistische Gewichtung	1.0	3424	14.1	.01	592	3424
gestutzter & kalibrierter Hochrechnungsfaktor	23.4	80000	47.4	.54	449	3424
gestutzte & kalibrierte Gewichtung	1.0	3424	2.0	.02	19	3424

1) Die tabellierten Verteilungskennwerte beziehenn sich auf die wellenspezifischen Samples an *teilnehmenden* Jugendlichen.

7 Forschungspraktische Hinweise

Zum Abschluss werden kurz einige Aspekte der Anwendung der entwickelten Gewichtungen im Rahmen von statistischen Analysen aufgegriffen. Dabei geht es zunächst um die Frage, *welche Gewichtung* für welche Analyse herangezogen werden soll. Daran schliessen einige Anmerkungen zur *Abschätzung* der auftretenden *Nonresponse-Verzerrungen* respektive zur Wirksamkeit der betreffenden in die Gewichtung eingebauten Korrekturen an. Abschliessend folgen einige Bemerkungen zur *Berechnung von Stichprobenfehlern* bzw. Signifikanztests auf der Basis des gewichteten Panelsamples.

Grundsätzlich wird für statistische Auswertungen stets die Gewichtung der *letzten* Befragungswelle herangezogen, anlässlich derer die Merkmale erhoben worden sind, die mit einer gegebenen Analyse ausgewertet werden. So werden beispielsweise die Gewichte der dritten Welle herangezogen, wenn PISA-Daten zusammen mit Merkmalen aus den ersten drei TREE-Wellen ausgewertet werden. Bei einer Querschnittanalyse anhand einer einzelnen Welle kommen die zu dieser Welle gehörenden Gewichte zum Einsatz. Diese Faustregel gilt unter der Annahme, dass für sämtliche Befragten Daten aus ein und denselben Wellen analysiert werden. Etwas anders verhält es sich, wenn beispielsweise ein Übertrittsprozess untersucht wird, der zu individuell variablen Zeitpunkten stattfindet und der entsprechend auch über unterschiedliche Befragungswellen erfasst wird (siehe z. B. Hupka, Sacchi & Stalder, 2006).

Die Einschlusswahrscheinlichkeit ins Analysesample und damit die Gewichtung (vgl. Beziehung 1) hängt in einem solchen Fall davon ab, mit welcher Welle der Übertritt erfasst worden ist, mit dem der untersuchte Lebensabschnitt endet. Den Jugendlichen wird für eine solche Analyse *individuell* die Gewichtung derjenigen Welle zugeordnet, mit der die benötigten Angaben zum Übertritt − oder allgemeiner: die zuletzt erhobenen der in einer Analyse benötigten Informationen − erfasst worden sind. Dafür werden die *rohen* Hochrechnungsgewichte im Datensatz herangezogen, da die kalibrierten Gewichte wegen der darin enthaltenen wellenspezifischen Rekalibrierungskonstanten nicht direkt über die Wellen vergleichbar sind. Ein mit Blick auf eine spezifische Analyse auf diese Weise neu zusammengestelltes rohes Hochrechnungsgewicht gilt es anschliessend wie beschrieben zu stutzen und zu kalibrieren; auf die Stutzung sollte dabei aus den erwähnten Gründen keinesfalls verzichtet werden (siehe Abschnitt 4). Dies gilt speziell, wenn ein Teil der analysierten Daten mit Welle 3 oder noch später erhoben worden ist. Das soweit beschriebene Vorgehen empfiehlt sich grundsätzlich immer dann, wenn die zuletzt erhobenen unter den für eine Analyse benötigten Daten von individuell unterschiedlichen Wellen stammen.

Ist die Gewichtung für eine gegebene Analyse ausgewählt respektive zusammengestellt und nötigenfalls rekalibriert (siehe Abschnitt 5) worden, so kann man sich die Frage stellen, wie sich die darin enthaltenen Nonresponse-Korrekturen auf die Schätzergebnisse auswirken. Um diese Frage zu beantworten, werden die entsprechend gewichteten Auswertungen am besten mit einer ansonsten identischen Analyse verglichen, der aber die − nötigenfalls entsprechend rekalibrierte − PISA-Basisgewichtung zugrunde gelegt wird. Stimmen die Ergebnisse in allen wesentlichen Punkten überein, so erlaubt dies den Schluss, dass die in der Panelgewichtung enthaltenen Nonresponse-Korrekturen die Schätzergebnisse nicht nennenswert beeinflussen. Andernfalls stellt sich die Frage, inwieweit die Unterschiede angesichts der Befunde aus Abschnitt 3 (insb. auch Abschnitt 3.11) wie auch vor dem Hintergrund des Forschungsstands zu Nonresponse und Stichprobenschwund plausibel sind. Der Effekt der PISA-Basisgewichtung auf die Schätzergebnisse ist dabei als weitgehend unproblematisch anzusehen, da er einzig auf das Erhebungsdesign von PISA 2000 zurückgeht (siehe Abschnitt 1). Die Unterschiede zwi-

schen Basis- und Panelgewichtungen beruhen dagegen auf den in den Abschnitten 3.2 bis 3.10 beschriebenen Schätzmodellen zur Korrektur von Nonresponseverzerrungen, womit neben Stichprobenfehlern stets auch mögliche Spezifikationsprobleme ins Spiel kommen (Menard 2002: 67f.). Dies gilt namentlich auch, weil bei der Konstruktion der Modelle jeweils relativ umfassend nach Untersuchungsmerkmalen gesucht worden ist, die die Teilnahme auf theoretisch plausible Weise beeinflussen. Ein solches teils induktives Vorgehen hat auf der einen Seite den Vorzug, dass allfällig vorliegende Nonresponseverzerrungen relativ umfassend korrigiert werden. Damit werden Spezifikationsfehler im Sinne eines *'Omitted Variable Bias'* (Menard 2002: 68f.) zumindest bezogen auf die empirisch erfassten Merkmale minimiert. Auf der anderen Seite birgt dieses Vorgehen aber unweigerlich auch die Gefahr eines *'Model-Overfit'* − also einer Überanpassung der Teilnahmemodelle ans spezifische Sample (siehe auch Wießner 2003: 89). Vor diesem Hintergrund ist es zu empfehlen, die Wirkung der Nonresponse-Korrektur auf die Schätzergebnisse wie eben beschrieben zu ermitteln und deren theoretische Plausibilität abzuschätzen.

Abschliessend soll noch kurz darauf hingewiesen werden, dass eine korrekte *Abschätzung der Samplefehler* mit dem gewichteten TREE-Panel die Berücksichtigung des komplexen Aufbaus der zu Grunde liegenden PISA-Ausgangsstichprobe erfordert. Auch wenn eine korrekt kalibrierte Panelgewichtung (siehe Abschnitt 6) verwendet wird, ermöglichen Statistikpakete, die, sei es explizit oder implizit, eine einfache Zufallsstichprobe voraussetzen, keine korrekte Schätzung der Samplevarianz. Diese werden *Standardfehler und Vertrauensintervalle* in der Regel vielmehr systematisch *unter-* und die *Signifikanzniveaus* entsprechend *überschätzen*. Stattdessen sollten für Varianzschätzungen mit dem TREE-Panel entweder induktive Replikations-Verfahren (z. B. Bootstrap-Methoden, vgl. Mooney & Duval 1993) oder aber spezialisierte Varianzschätzer für komplexe Stichproben herangezogen werden (siehe Lee, Forthofer & Lorimor 1989), wie sie etwa in STATA oder in neueren Versionen von SPSS implementiert sind (STATA: 'svy'-Befehle; SPSS: 'complex samples'-Tools).[1]

[1] Die dafür erforderlichen Annahmen, die Konstruktion der benötigten Hilfsvariablen und einige Anwendungsbeispiele sind einem separaten Dokument zusammengestellt (Sacchi 2008). Vgl. auch Codebook zum Gewichtungsdatensatz des jeweils jüngsten TREE Data Releases.

Literaturhinweise

Elliot, Dave (1991). 'Weighting for Non-Response. A Survey Researcher's Guide'. Office of Population Censuses and Surveys (OCPS), Social Survey Division: London.

Everitt, Brian S. (1993). 'Cluster Analysis'. Edward Arnold: London (3. Auflage).

Hagenaars, Jacques A. (1990). 'Categorical Longitudinal Data. Log-Linear Panel, Trend, and Cohort Analysis'. Sage: Newbury Park.

Haisken-DeNew, John & Joachim Frick (2000). 'Desktop Companion to the German Socio-Economic Panel (GSOEP)'. DIW Berlin (4. Auflage).

Hosmer, David W. & Stanley Lemeshow (1989). 'Applied Logistic Regression'. John Wiley & Sons: New York.

Hupka, Sandra, Stefan Sacchi & Barbara Stalder (2006). 'Herkunft oder Leistung? Analyse des Eintritts in eine zertifizierende nachobligatorische Ausbildung anhand des Jugendlängsschnitts TREE'. TREE: Bern.

Jaccard, James (2001). 'Interaction Effects in Logistic Regression'. Bd. 135 der Reihe 'Sage University Paper Series on Quantitative Applications in the Social Science', hg. von Michael S. Lewis-Beck. Sage: Beverly Hills.

Kish, Leslie (1992). 'Weighting for Unequal Pi', Journal of Official Statistics, Vol. 8 (2): 183-200.

Kish, Leslie (1995 [1965]). 'Survey Sampling'. John Wiley: New York.

Klein, Sabine & Rolf Porst (2000). 'Mail Surveys. Ein Literaturbericht'. Technischer Bericht 10. ZUMA.

Koch, Achim & Rolf Porst (Hg.) (1998). 'Nonresponse in Survey Research'. Reihe 'Proceedings of the Eighth International Workshop on Houshold Survey Nonresponse' ZUMA: Mannheim.

Lee, Eun Sul, Ronald N. Forthofer & Ronald J. Lorimor (1989). 'Analyzing Complex Survey Data'. Bd. 71 der Reihe 'Quantitative Applications in the Social Science', hg. von Michael Lewis-Beck. Sage: Newbury Park.

Menard, Scott (2002). 'Applied Logistic Regression'. Bd. 76 der Reihe 'Sage University Paper Series on Quantitative Applications in the Social Sciences', hg. von Michael S. Lewis-Beck. Sage: Thousand Oaks (2. Auflage).

Meyer, Thomas (2000). 'Evaluation des TREE-Adressblätter-Rücklaufs'. Internes Arbeitspapier. TREE: Bern.

Mooney, Christopher Z. & Robert D. Duval (1993). 'Bootstrapping. A Nonparametric Approach to Statistical Inference'. Bd. 95 der Reihe 'Quantitative Applications in the Social Science', hg. von Michael S. Lewis-Beck. Sage: Newbury Park.

Moser, Claus A. & Graham Kalton (1971). 'Survey Methods in Social Investigation'. Heinemann: London (2. Edition).

PISA Consortium (2000a). 'PISA International Data Base'. OECD (Ed.).

PISA Consortium (2000b). 'PISA Weighting and Variance Estimation'. OECD (Ed.).

PISA, Programme for International Student Assessment (2002). 'Für das Leben gerüstet? Die Grundkompetenzen der Jugendlichen − Nationaler Bericht der Erhebung PISA 2000'. Bundesamt für Statistik, EDK: Neuchâtel.

PISA Romandie (ohne Jahr). 'La pondération de l'échantillon des élèves de 9e pour l'enquête PISA en Suisse romande'.

Renaud, Anne, Erich Ramseier & Claudia Zahner (2000). 'PISA 2000: Sampling in Switzerland. General Information and Design'. PISA.ch (Hg.): 'Report for the International Consortium'.

Rizzo, Lou, Graham Kalton & J. Michael Brick (1994). 'Weighting Adjustments for Panel Nonresponse in the SIPP'. Final Report. Westat Inc.: Rockville.

Sacchi, Stefan (2001). 'Longitudinal-Gewichtungen ausgewählter Haushaltspanels. Review im Auftrag des schweizerischen Haushaltspanels'. Cue Sozialforschung: Zürich.

Sacchi, Stefan (2003). 'Longitudinale Stichprobengewichtung fur das TREE-Panel (Befragungswellen 1 & 2)'. Cue Sozialforschung: Zürich.

Sacchi, Stefan (2004a). 'Revision der longitudinalen Stichprobengewichtung für das TREE-Panel (Befragungswellen 1 & 2)'. Cue Sozialforschung: Zürich.

Sacchi, Stefan (2004b). ‘Longitudinale Stichprobengewichtung für Welle 3 des TREE-Panels’. Cue Sozialforschung: Zürich.

Sacchi, Stefan (2008a). TREE-Längsschnittgewichtung: Konstruktion und Anwendung. Dokumentation zu den acht Erhebungswellen 2000 bis 2007. TREE und cue sozialforschung: Bern/Zürich.

Sacchi, Stefan (2008b). ‘Varianzschätzung mit dem TREE-Panel’. Cue Sozialforschung: Zürich.

Stalder, Barbara & Dellenbach, Myriam (2005). Dokumentation der Non-Response-Befragungen im Anschluss an die TREE-Erhebungswellen 2003 und 2004. Internes Arbeitspapier. Bern: TREE.

Scherpenzeel, Annette (2001). ‘Mode Effects in Panel Surveys: A Comparison of CAPI and CATI’. Bundesamt für Statistik (Hg.): ‘BFS Aktuell’: Neuchâtel.

Schnell, Rainer (1997). ‘Nonresponse in Bevölkerungsumfragen. Ausmass, Entwicklung und Ursachen’. Leske + Budrich: Opladen.

Wießner, Frank (2003). 'Nonresponse bei Verbleibsuntersuchungen. Korrekturverfahren zu Antwortausfällen am Beispiel ehemals arbeitsloser Existenzgründer, die mit dem Überbrückungsgeld (§57 SGB III) gefördert wurden.' Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 36 (1): 77-96.